Panel data econometrics and climate change

Muris, C.H.M.

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

Panel data econometrics and climate change

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit van Tilburg op gezag van de rector magnificus, prof. dr. Ph. Eijlander, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op vrijdag 1 april 2011 om 14.15 uur door

Christiaan Hubertus Maria Müris

geboren op 18 maart 1984 te Venlo.

Promotor:	prof. dr. J.R. Magnus
Promotiecommissie:	prof. dr. B. Melenberg
	prof. dr. J.A. Smulders
	prof. dr. M. Verbeek
	dr. R.J.A. Laeven
	dr. M. Ikefuji
	dr. M. Wild

Veur Djavon

Acknowledgements

Let me start by thanking Jan Magnus for being a great advisor. If this thesis is a success and/or my career turns out to be a success, that is thanks to him. I am deeply grateful for everything he has taught me, and for everything he has done for me.

Bertrand Melenberg has been teaching and mentoring me since my second year of undergraduate studies. He has helped me complete a Bachelor, Master, and this PhD thesis. He has greatly influenced the way I think about many topics, including, but not limited to, econometrics. He has made my stay in Tilburg more pleasant and more productive.

Roger Laeven and Masako Ikefuji have been wonderful coauthors, and I thank them and Sjak Smulders, Marno Verbeek, and Martin Wild for agreeing to be on my PhD committee.

I want to thank my parents for all their love and care, and for having supported me regardless of what I chose to do. As ik eine kleine heb, probaer ik ut precies zoe te doon wie og. Finally, Andrea: thank you for everything.

Contents

Chapter 1

Introduction and summary

This thesis consists of three parts. The first part (Chapter 2) applies methods from *panel data econometrics* to a problem in *climate change*. The second part (Chapters 3, 4, and 5) is concerned with an economic model of *climate change*. The third part (Chapters 6 and 7) is a contribution to the theory of *panel data* methods. The chapters are based on the following research papers:

- *Chapter 2:* Magnus, J.R., B. Melenberg, and C. Muris (2010), Global warming and local dimming: the statistical evidence, forthcoming in the *Journal of the American Statistical Association.*
- *Chapter 3:* Ikefuji, M., R.J.A. Laeven, J.R. Magnus, and C. Muris (2010), Expected utility and catastrophic risk in a stochastic economy-climate model, *Working paper.*
- *Chapter 4:* Ikefuji, M., R.J.A. Laeven, J.R. Magnus, and C. Muris (2010), Burr utility, *CentER Discussion Paper*, 2010-81, Tilburg University.
- *Chapter 5:* Ikefuji, M., R.J.A. Laeven, J.R. Magnus, and C. Muris (2010), Scrap value functions in dynamic decision problems, *CentER Discussion Paper*, 2010-77, Tilburg University.
- *Chapter 6:* Magnus, J.R., and C. Muris (2010), Specifications of variance matrices for panel data models, *Econometric Theory*, 26, 301–310.
- *Chapter 7:* Muris, C. (2010), Efficient GMM estimation with a general missing data pattern, *Working paper.*

In Chapter 2, we investigate the relative importance of two opposing effects on global temperature. On one hand, there is the greenhouse effect: an increase in concentrations of carbon dioxide and other greenhouse gases warms the planet. On the other hand, there is the solar radiation effect: aerosols reflect and absorb sunlight in the atmosphere so that less sunlight reaches the Earth, so the Earth becomes cooler. Decomposing the two effects is important because the existence of the solar radiation effect obscures the magnitude of the greenhouse effect. Identifying the two effects is not straightforward because we only observe the sum of the two. We manage to overcome this obstacle by using a simple climate model, and weather station data for the period 1959–2002. We find that the estimated global temperature change of 0.73 °C can be decomposed into a greenhouse effect of 1.87 °C, a solar radiation effect of −1.09 °C, and a small remainder term.

In Chapter 3, we show that an economic model of climate is fragile when introducing heavy-tailed uncertainty. Both the way that we model the economy and the way we introduce uncertainty are according to standard practice, suggesting that the tools used by economists and policy advisors are subject to this fragility. We derive necessary and sufficient conditions on the economic model to avoid this fragility, and then solve our model for two examples of non-fragile utility functions. We also develop and implement a procedure to learn the input parameters of our model and show that the model thus specified produces non-fragile optimal policies. Chapters 4 and 5 address methodological issues that relate to the analysis in Chapter 3.

In Chapter 6, we consider the variance matrix in panel data models. The dimension of this matrix is $TN \times TN$, where N is the number of countires, households, or firms, and T is the number of time periods. If TN is large, working with this variance matrix is not practical from a computational point of view. We define structures for the variance matrix that allow the computation of the essential quantities using matrices of sizes T and N only. In these cases, working with the matrix is easy, while retaining flexibility of the panel data model. In particular, we allow for heteroskedasticity, for household- or station-specific correlation, and for time-specific spatial correlation.

Chapter 7 is concerned with missing data. I propose an estimator that efficiently uses all the available information in the data set, even when some of the observations are incomplete. The estimator can be applied to a wide variety of econometric models, and does not restrict the missing data patterns found in the data. This makes the estimator very flexible. At the same time, it is easy to implement and computationally cheap.

Chapter 2

Global warming and local dimming: the statistical evidence

Abstract: Two effects largely determine global warming: the well-known greenhouse effect and the less well-known solar radiation effect. An increase in concentrations of carbon dioxide and other greenhouse gases contributes to global warming: the greenhouse effect. In addition, small particles, called aerosols, reflect and absorb sunlight in the atmosphere. More pollution causes an increase in aerosols, so that less sunlight reaches the Earth (global dimming). Despite its name, global dimming is primarily a local (or regional) effect. Because of the dimming the Earth becomes cooler: the solar radiation effect. Global warming thus consists of two components: the (global) greenhouse effect and the (local) solar radiation effect, which work in opposite directions. Only the sum of the greenhouse effect and the solar radiation effect is observed, not the two effects separately. Our purpose is to identify the two effects. This is important, because the existence of the solar radiation effect obscures the magnitude of the greenhouse effect. We propose a simple climate model with a small number of parameters. We gather data from a large number of weather stations around the world for the period 1959–2002. We then estimate the parameters using dynamic panel data methods, and quantify

the parameter uncertainty. Next, we decompose the estimated temperature change of 0.73 °C (averaged over the weather stations) into a greenhouse effect of 1.87 °C, a solar radiation effect of −1.09 °C, and a small remainder term. Finally, we subject our findings to extensive sensitivity analyses.

2.1 Introduction

The Earth is getting warmer and much or all of this process is generally believed to be caused by humans. There is much uncertainty about global warming. The purpose of this paper is to investigate the statistical evidence of global warming, using econometric panel data techniques supplemented by extensive sensitivity analyses.

We distinguish between two effects which together largely determine global warming. First, the concentrations of carbon dioxide (CO2) and other 'greenhouse gases' have increased. For example, the amount of CO2 in the atmosphere has increased by about 36% between 1750 and 2005 (Solomon et al, 2007, Chapter 2, p. 137). These greenhouse gases act as a blanket, thus contributing to global warming: the greenhouse effect. Because of the long lifetime of CO2 in the atmosphere, this effect is global.

The second effect, not as well known by the general public, is the solar radiation effect. Pollution consists, in part, of small particles, called 'aerosols', which reflect and absorb sunlight in the atmosphere and make clouds more reflective. More aerosols implies that less sunlight reaches the Earth: global dimming (Power, 2003; Norris and Wild, 2007; Wild, 2009). Global dimming varies in time and location. The term 'global' in 'global dimming' is somewhat misleading, because it refers to the sum of diffuse and direct solar radiation (global radiation), and not to a global scale of the phenomenon (Wild, 2009, p. 1). In fact, dimming is primarily a local or regional effect, because aerosols have a short lifetime (about one week) in contrast to greenhouse gases which have a lifetime of up to 100 years (Kaufman et al, 2002). As a result of the dimming the Earth becomes cooler: the solar radiation effect (Haywood and Boucher, 2000; Ramanathan et al, 2001; Kaufman et al, 2002; Bellouin

et al, 2005). Global warming thus consists of two components: the (global) greenhouse effect and the (local) solar radiation effect, which work in opposite directions.

When we observe an increase in temperature, we observe only the sum of the greenhouse effect and the solar radiation effect, but not the two effects separately. Our purpose is to try and identify the two effects. This is important because policy makers are successful in reducing aerosols (which has a local benefit) but less successful in reducing CO2 (which has a global, but almost no local benefit). A reduction in aerosols causes cleaner air (good), but also more solar radiation (bad). The solar radiation effect thus obscures the magnitude of the greenhouse effect, and forecasts ignoring the solar radiation effect underestimate the increase in temperature. The size of the solar radiation effect is uncertain (Anderson et al, 2003; Andreae et al, 2005), and hence the solar radiation effect offsets the greenhouse effect by an unknown amount.

Current methods to assess the effect of greenhouse gases in the presence of aerosols typically use global climate models, requiring a large number of parameters whose values are typically obtained by calibration rather than estimation. The reliability of such models is reviewed in Räisänen (2007). The values for the effect of greenhouse gases and aerosols on temperature vary greatly (Anderson et al, 2003; Roe and Baker, 2007), thus adding to the controversy about climate change.

Our approach is different. We propose a simple climate model with a small number of parameters. We gather data from a large number of weather stations around the world for the period 1959–2002. We estimate the parameters using dynamic panel data methods, and quantify the parameter uncertainty. Then we decompose the observed temperature change into a greenhouse and a solar radiation effect.

This paper is organized as follows. In Section 2.2 we discuss the energy balance, which is used to construct our climate model. In Section 2.3, we describe our datasources, the construction of our dataset, and how we have dealt with a selection problem. The econometric model is presented in Section 2.4. We report our results and the decomposition in greenhouse and solar

radiation effects in Section 2.5, and we offer extensive sensitivity analyses in Section 2.6. Section 2.7 concludes.

2.2 The energy balance

The Earth and its atmosphere receive energy from the Sun in the form of shortwave radiation, which is partly absorbed, and the energy associated with the absorbed radiation is returned to space as longwave radiation. As long as the amount of incoming solar radiation absorbed by Earth and atmosphere is balanced by Earth and atmosphere releasing the same amount of outgoing radiation, the Earth's temperature will remain the same. A simplified scheme of the energy balance is given in Figure 2.1, which is based on Trenberth et al (2009); see also McGuffie and Henderson-Sellers (2001).

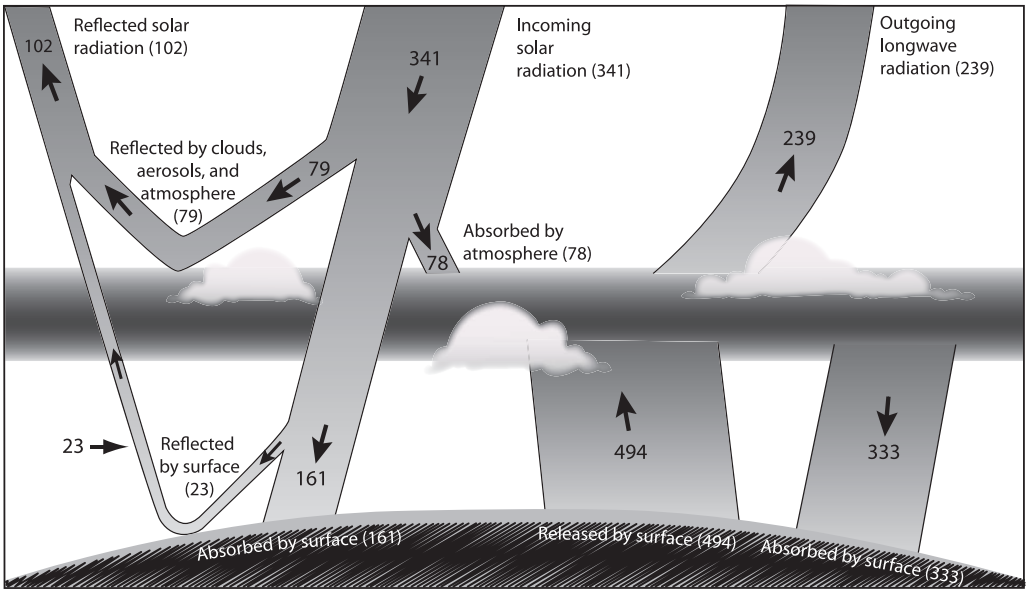


Figure 2.1: The Earth's annual energy balance (Wm^{-2}). Adapted from Trenberth et al (2009).

The amount of solar radiation reaching the Earth's atmosphere is about 341 Watts per meter squared (Wm^{-2}). Solar radiation has a short wavelength, and hence most of the solar radiation passes through the atmosphere

and reaches the surface of the Earth (184 Wm^{-2}). Some of the solar radiation, however, is reflected back into space (79 Wm^{-2}) due to clouds and small particles (aerosols) in the atmosphere, and some is absorbed (78 Wm^{-2}) in the atmosphere where it is transferred to heat energy and longwave radiation. When the Sun's radiation reaches the Earth, part is absorbed (161 Wm^{-2}) and transferred to longwave radiation, and part is reflected back into space as shortwave radiation (23 Wm^{-2}). The Earth releases energy (494 Wm^{-2}), consisting of longwave radiation (396 Wm^{-2}) and latent and sensible heat (98 Wm^{-2}). Most of the emitted longwave radiation is absorbed in the atmosphere by clouds and so-called greenhouse gases. The longwave radiation emitted by the atmosphere goes back into space (239 Wm^{-2}) or is radiated back to Earth (333 Wm^{-2}).

The energy absorbed by the Earth's surface thus consists of two components: shortwave from the Sun (161 Wm^{-2}) and longwave from the atmosphere (333 Wm^{-2}). Without the longwave component the average temperature on Earth would be about -18 °C, while in fact it is about 13.5 °C. The longwave component exists because of the presence of greenhouse gases (and clouds), which act as a blanket for the longwave radiation coming from the Earth's surface (McGuffie and Henderson-Sellers, 2001): the *greenhouse effect*. One of the most important greenhouse gases is carbon dioxide (CO2). While the natural greenhouse effect is crucial for the climate on Earth, human activities have intensified it. For example, the amount of CO2 in the atmosphere has increased by about 36% between 1750 and 2005, primarily through the combustion of fossil fuels and tropical deforestation, and by about 15% between 1975 and 2005; see Solomon et al (2007, Chapter 2, p. 137). The Earth becomes warmer (*global warming*) and the anthropogenic greenhouse effect is thought to be primarily responsible for the speed at which this happens (Solomon et al, 2007, Chapter 9, p. 665). The greenhouse effect is a global effect, and hence heavy industries and deforestation in one area affect people everywhere.

Increased pollution not only results in a higher concentration of CO2, but also in more aerosols. An increase in aerosols implies that less sunlight

reaches the Earth's surface (*global dimming*), and hence that the Earth becomes cooler: the *solar radiation effect*. Global warming thus consists of two components: the greenhouse effect and the solar radiation effect, which work in opposite directions.

We propose a climate model based on the simplified energy balance described above. Our model is inspired by the energy balance models proposed by Budyko (1969), Sellers (1969), North et al (1981), and others; see also Gregory et al (2002), Andreae et al (2005), and Schwartz (2007) for recent applications.

If the energy balance at the Earth would hold exactly, then (combining the energy balances at the Earth's surface and the atmosphere)

$$E^{sin} - E^{lout} = 0, \tag{2.1}$$

where $E^{sin} = (161 + 78)$ Wm^{-2} denotes the incoming solar shortwave radiation which reaches and is absorbed by the Earth or the atmosphere and $E^{lout} = 239$ Wm^{-2} is the longwave radiation emitted from the atmosphere. In reality, the energy balance will not hold exactly and this imbalance will result in a change in temperature, modeled as

$$\frac{c\,(\mathrm{TEMP}_{t+\Delta t} - \mathrm{TEMP}_t)}{\Delta t} = E_t^{sin} - E_t^{lout}, \tag{2.2}$$

where c is the so-called 'heat capacity', linking the energy surplus or deficit to a change in temperature per unit of time (Andreae et al, 2005).

While Equations (2.1) and (2.2) refer to the Earth as a whole, we wish to consider weather stations on the Earth's surface. The energy balance (2.1) then still applies with two modifications. First, the various energy terms will be station-specific. Second, weather stations near the equator (latitude zero) receive more sunlight than stations at lower or higher latitudes. Some of this excess radiation will flow from warmer areas to colder areas, resulting in an additional term E^{exch}, representing the net in- or outflow of energy. Thus, if the energy balance would hold exactly in weather station i, then

$E_{it}^{sin} - E_{it}^{lout} + E_{it}^{exch} = 0$, but when there is an imbalance, the discrepancy will result again in a change in local temperature TEMP_{it}, modeled for station i at time t as

$$\frac{c\,(\text{TEMP}_{i,t+\Delta t} - \text{TEMP}_{it})}{\Delta t} = E_{it}^{sin} - E_{it}^{lout} + E_{it}^{exch}. \tag{2.3}$$

Equation (2.3) is the starting point for our econometric climate model. The four energy terms will depend on solar radiation, greenhouse gas concentration, and temperature.

2.3 Data and descriptive statistics

We require annual data at the level of weather stations. For each station we collected monthly observations on temperature (TEMP): the average temperature in degrees Celsius (°C) at the surface (source: CRU); solar radiation (RAD): the amount of sunlight ('global solar irradiance') that reaches the Earth's surface, measured in Watts per meter squared (Wm^{-2}) (source: GEBA); and carbon dioxide (CO2): concentration of carbon dioxide, measured in parts per million by volume (ppmv) (source: Mauna Loa Observatory). In addition, we need for each station its longitude and latitude. The data are constructed from three sources.

The *Climatic Research Unit (CRU)* maintains a database of monthly climate observations based on a large number of weather stations around the globe (land stations only, Antarctica excluded) over the period January 1901 to December 2002. We use the database labeled CRU TS 2.1. The database can be found online at `http://www.cru.uea.ac.uk`. Information is provided on nine climate variables including TEMP. Some areas of the Earth contain more weather stations than others. In order to obtain regularity of information, the surface of the Earth is defined on a high-density (0.5°) latitude-longitude grid, thus dividing the Earth in 720×360 grid cells, each covering an area of about 45×45 kilometers. Each grid cell draws potential information from about 100 weather stations, both within and in the neighborhood

of the grid cell. The landmass (excluding Antarctica) covers about 26.5% of the Earth. Monthly information is thus provided for each of the nine climate variables in each of 67,420 cells on the landmass. The construction of the database includes checks for inhomogeneities, the use of neighboring stations to fill in gaps, and spatial and temporal interpolation using station data from different datasets (Mitchell and Jones, 2005). There exist other sources for TEMP, such as the weather station data from the National Climatic Data Center (NCDC). The CRU dataset is, however, the most extensive, and where the CRU and NCDC data overlap geographically we do not find systematic differences.

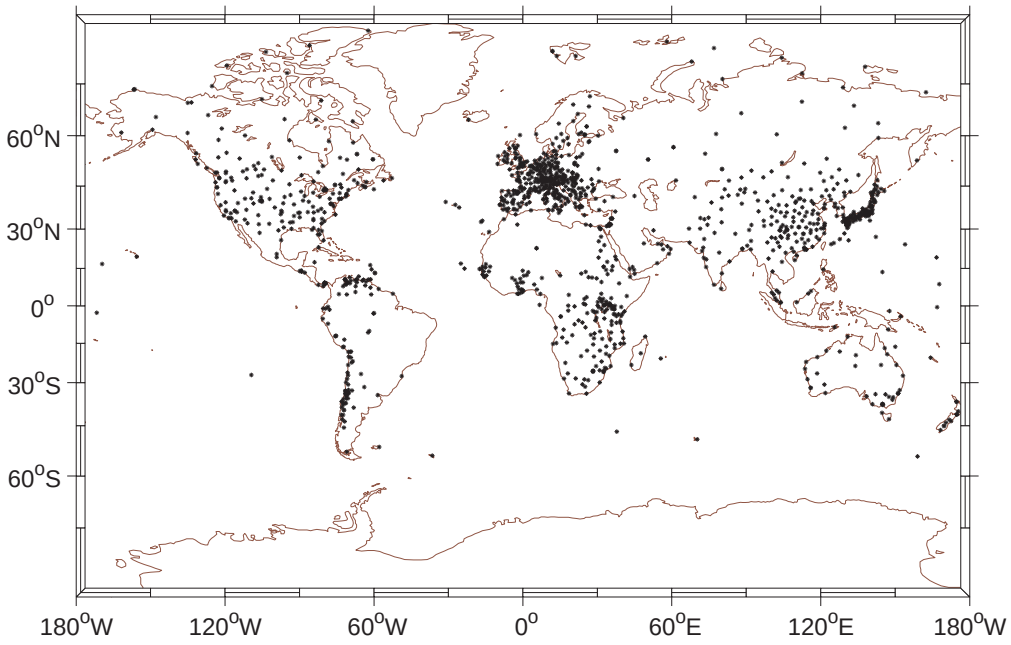


Figure 2.2: Distribution of weather stations in the GEBA dataset

The *Global Energy Balance Archive (GEBA)* is project A7 of the World Climate Programme—Water (WMO/ICSU). The GEBA database stores monthly means of energy fluxes which have been instrumentally measured at the surface, and is publicly available (http://bsrn.ethz.ch/gebastatus). The quality of the energy flux monthly means is controlled. The database provides us with monthly observations on solar radiation over the period 1950–2006, under both cloudy and cloudfree conditions. We only consider the observations

from January 1959 to December 2002, because the CO2 data are not available before 1959 and the CRU data are not available after 2002. Over this 44-year period the GEBA database contains monthly data from 2164 weather stations around the Earth. We delete stations on boats and stations with a quality flag (unreliable). Of the remaining stations there are many where some of the observations are missing. We include only those stations which have at least one complete year of observations. This leaves us with 1337 stations. Figure 2.2 shows that the weather stations are not spread evenly over the continents, and this could have implications which we discuss and resolve in Section 2.6. If the solar radiation data on these 1337 stations were complete we would have $44 \times 1337 = 58,828$ complete years, while in fact we have only $18,604$ complete years. An average weather station has thus only about fourteen complete years of solar radiation data. The 'holes' can occur at the beginning, the middle, or the end of each time series. For the GEBA weather stations the geographical information on longitude and latitude (and elevation) is also available. See Gilgen and Ohmura (1999) for a detailed description of the GEBA database.

The *Mauna Loa Observatory (MLO)* in Hawaii is one of the baseline observatories of the National Oceanic and Atmospheric Administration. The dataset we are using is the oldest continuous carbon dioxide concentration dataset available, and provides monthly and annual data on CO2, the concentration of carbon dioxide, measured in parts per million volume, from January 1959 to the present. It is publicly available (http://www.mlo.noaa.gov/home.html). Since CO2 is well-mixed in the atmosphere (Solomon et al, 2007, Chapter 2, p. 138), we may assume that CO2 is the same for each weather station and hence we don't require CO2 data at station level.

From these three sources we obtain monthly observations on TEMP (1901–2002); RAD and geographical variables (1950–2006); and CO2 (1959–present). This gives a period of 44 years (1959–2002) for which all variables are observed. To construct a consistent dataset over the 1959–2002 period we add TEMP to the RAD dataset. Given the location of the weather stations in the RAD dataset, and the division of the Earth into grid cells by CRU, we determine for each RAD station the corresponding grid cell in the CRU division, and thus allocate to each RAD station the appropriate CRU data. We use annual data

rather than monthly data in order to avoid the difficult problem of seasonal adjustments. The annual data are obtained by simple averaging of the monthly data, except for the CO2 series where annual data are provided by the Mauna Loa Observatory. This results in a panel dataset consisting of observations over 1337 weather stations during 44 years.

Monthly observations on TEMP are available, but only about 32% of the monthly observations on RAD is available. When solar radiation is not observed at some weather station during one of the months in a particular year, the corresponding observation is classified as a *missing item observation* (where 'missing item' applies to missing information on solar radiation only). As a consequence our dataset is an unbalanced panel with 18,604 (out of a possible 58,828) annual observations without missing items.

Variable		Mean	Std.	Min	Max
TEMP	overall	13.40	8.90	−22.04	31.23
complete panel	between		8.89	−19.96	29.75
	within		0.34	12.91	14.14
TEMP	overall	11.93	8.43	−22.04	30.36
unbalanced panel	between		8.90	−20.74	29.77
	within		0.61	10.66	13.21
RAD	overall	160.91	42.46	52.00	324.00
unbalanced panel	between		44.68	55.46	316.00
	within		9.09	148.77	183.21
CO2		340.88	17.55	315.98	373.10

Table 2.1: Sample statistics for TEMP, RAD, and CO2

Table 2.1 presents the sample statistics for TEMP, RAD, and CO2. For temperature we present information both for the 'complete panel' (the panel including the missing item observations) and for the 'unbalanced panel' (the panel without the missing item observations). For solar radiation we can only present information for the unbalanced panel, and for CO2 we present the sample statistics based on the annual data. The rows labeled 'overall' consider all the data (58,828 for TEMP in the complete panel, 18,604 for TEMP and RAD in the unbalanced panel, and 44 for CO2). The rows labeled 'between' consider cross-section averages (1337 stations), and the rows labeled 'within'

consider time-series averages (44 years for the complete panel and 13.91 years for the unbalanced panel.) We see from Table 2.1 that the sample average of solar radiation in the unbalanced panel is 160.91 Wm^{-2}, ranging from a lowest year average (over weather stations) of 148.77 Wm^{-2} to a highest year average of 183.21 Wm^{-2}, and that the level of CO2 at the Mauna Loa Observatory increased from 315.98 ppmv in 1959, the first year of the panel, to 373.10 ppmv in 2002, the final year.

Variable		Mean	Std.	Min	Max
ΔTEMP	overall	0.0142	0.7311	−4.9250	5.1583
complete panel	between		0.0162	−0.0583	0.0944
	within		0.2580	−0.5140	0.5726
ΔTEMP	overall	0.0136	0.7600	−4.9250	5.1583
unbalanced panel	between		0.2802	−3.6167	1.5500
	within		0.3451	−0.6495	0.8305

Table 2.2: Sample statistics for time differences in temperature

The average temperature in the complete panel is 13.4 °C, ranging from a year average (over all weather stations) of 12.91 °C in the coldest year to 14.14 °C in the warmest year, and ranging from a station average (over all years) of −19.96 °C in the coldest weather station to 29.75 °C in the warmest weather station. In the unbalanced panel some of the temperature averages are substantially lower, up to almost 1.5 °C. This suggests that the missing observations may not be missing completely at random (MCAR), and hence that a (potentially serious) sample selection problem may exist, at least in terms of the *level* of temperature. We are, however, primarily interested in a decomposition of temperature *changes* (in the time period 1959 to 2002). To investigate whether there is a selection problem due to missing item observations in terms of temperature changes we present in Table 2.2 the complete and unbalanced panel for time differences in temperature. Because we take first differences there are now only 43 years and hence $43 \times 13,337 = 57,491$ observations for TEMP in the complete panel, and 15,388 in the unbalanced panel. The average annual temperature change in the complete panel is 0.0142 °C, only slightly higher than the average annual temperature change in the unbalanced panel (0.0136 °C). The overall difference between the two panels is

thus only 0.0006 °C per year, and this difference is statistically not significant (p-value = 0.85). For individual weather stations the time averages in the complete and unbalanced panels sometimes differ substantially. This is because for some weather stations only a few years are without missing items, implying that extreme weather conditions may have a large impact for these stations. This is also reflected by the corresponding 'between' standard deviations: only 0.0162 in the complete panel, but 0.2802 in the unbalanced panel.

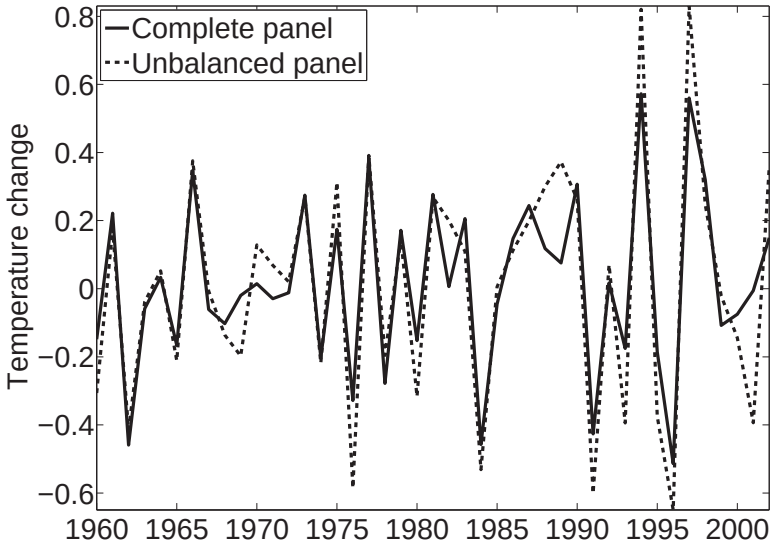


Figure 2.3: Average temperature change, 1960–2002

Regarding the year averages over weather stations (the two rows labeled 'within'), we see that the difference between the complete and unbalanced panel is small, and this is further illustrated in Figure 2.3, where we present the annual temperature changes (averaged over all weather stations) in both the complete and the unbalanced panel for 1960–2002. We tested the null hypothesis that the mean temperature changes for each of the years from 1960 to 2002 in both panels are equal, but could not reject the null hypothesis (p-value = 1.00). Hence we conclude that, when dealing with temperature changes, we may treat the missing observations as MCAR.

The average temperature change over the weather stations in our panel is not necessarily the same as the 'global' average temperature change. However,

a comparison of our average temperature change with the 'global' average temperature change based on the CRU data for land air temperature or the CRU data for combined land and marine temperature, indicates that the decomposition of our average temperature change (into the greenhouse and radiation effects) will also be informative for these 'global' temperature changes.

2.4 The econometric model

2.4.1 Specification of the energy flows

Our econometric model is based on Equation (2.3) in annual terms ($\Delta t = 1$ year):

$$c\,(\mathrm{TEMP}_{i,t+1} - \mathrm{TEMP}_{it}) = E_{it}^{sin} - E_{it}^{lout} + E_{it}^{exch}, \tag{2.4}$$

where the energy terms represent annual measurements. Let us specify the three energy flows, following Budyko (1969) with minor modifications; see also Sellers (1969), North (1975), and North et al (1981).

We allow for both a global and a local solar radiation effect, and we therefore specify

$$E_{it}^{sin} = a_0 + a_1\overline{\mathrm{RAD}}_t + a_2(\mathrm{RAD}_{it} - \overline{\mathrm{RAD}}_t),$$

where $\overline{\mathrm{RAD}}_t$ denotes the average solar radiation at year t and $(\mathrm{RAD}_{it} - \overline{\mathrm{RAD}}_t)$ the local solar radiation in excess of average solar radiation. We have $a_1 \geqslant a_2 \geqslant 0$, because an increase in either $\overline{\mathrm{RAD}}_t$ or RAD_{it} leads to an increase in E_{it}^{sin}. The global effect is captured by $a_1\overline{\mathrm{RAD}}_t$, while $a_2(\mathrm{RAD}_{it} - \overline{\mathrm{RAD}}_t)$ captures the local effect. There is no global effect if $a_1 = a_2$, and no local effect if $a_2 = 0$. We shall assume that changes in solar radiation are caused by changes in anthropogenic aerosol emissions: more aerosols lead to a decrease in solar radiation (Power, 2003; Norris and Wild, 2007). Our analysis does not, however, depend on this assumption, and changes in solar radiation can also be influenced by other factors, such as variations in the solar constant.

The outgoing longwave energy is an increasing (nonlinear) function of temperature, and also depends on the concentration of greenhouse gases in the

atmosphere, which we represent by the concentration of CO2. Assuming a constant vertical lapse rate (cf. North, 1975), the atmosphere's temperature depends linearly on the Earth's surface temperature. Since greenhouse gases are assumed to be evenly spread around the globe, we model their effect to be constant over weather stations. Based on these considerations, we approximate the outgoing longwave energy by the following linear function:

$$E_{it}^{lout} = b_0 + b_1\overline{\mathrm{TEMP}}_t + b_2(\mathrm{TEMP}_{it} - \overline{\mathrm{TEMP}}_t) - b_3\log(\mathrm{CO2}_t),$$

where $\overline{\mathrm{TEMP}}_t$ denotes the average temperature at year t, $b_1 \geqslant b_2 \geqslant 0$, and $b_3 \geqslant 0$. Again, we allow for both a local and a global effect. Finally, the exchange energy term is modeled as

$$E_{it}^{exch} = c_0 - c_1\left(\mathrm{TEMP}_{it} - \overline{\mathrm{TEMP}}_t\right)$$

with $c_1 \geqslant 0$. Thus, if the local temperature in weather station i is larger than the average temperature, then there is an outflow of energy from station i; if the local temperature is lower than the average, there is an inflow. The parametrizations for E_{it}^{lout} and E_{it}^{exch} are based on Budyko (1969), North (1975), and North et al (1981). The dependence on CO2 via a log-transformation is based on Solomon et al (2007, Chapter 2, p. 140).

With these specifications substituted into Equation (2.4) we obtain, after suitable parameter transformations,

$$\mathrm{TEMP}_{i,t+1} = \beta_1\mathrm{TEMP}_{it} + \beta_2\mathrm{RAD}_{it} + \lambda_t, \tag{2.5}$$

$$\lambda_t = \gamma_0 + \gamma_1\overline{\mathrm{TEMP}}_t + \gamma_2\overline{\mathrm{RAD}}_t + \gamma_3\log(\mathrm{CO2}_t). \tag{2.6}$$

We can estimate the β's and the γ's, but not the underlying structural parameters, unless we make further assumptions, for example, about the heat capacity c.

2.4.2 Steady state

The system gives rise to a steady state temperature, both at a global and at a local level, obtained by setting $\mathrm{TEMP}_{i,t+1} = \mathrm{TEMP}_{it}$ for all weather stations i at a given year t. The global average steady state temperature at year t will be denoted by $\overline{\mathrm{TEMP}}_t^e$ and the local steady state temperature in weather station i at year t by TEMP_{it}^e. The steady state temperatures are then given by

$$\overline{\mathrm{TEMP}}_t^e = \frac{\gamma_0 + (\beta_2 + \gamma_2)\overline{\mathrm{RAD}}_t + \gamma_3 \log(\mathrm{CO2}_t)}{1 - \beta_1 - \gamma_1} \tag{2.7}$$

and

$$\mathrm{TEMP}_{it}^e = \overline{\mathrm{TEMP}}_t^e + \frac{\beta_2}{1 - \beta_1}\left(\mathrm{RAD}_{it} - \overline{\mathrm{RAD}}_t\right). \tag{2.8}$$

The global average steady state temperature is thus determined by the global average solar radiation level and the level of the greenhouse gases (represented by CO2). The local steady state temperature may deviate from the global average steady state temperature via a deviating local solar radiation level.

Using the steady state temperatures (2.7) and (2.8) we can decompose a change in local or global steady state temperature into a solar radiation effect and a greenhouse effect. For example, a change in global steady state temperature is given by

$$\Delta\overline{\mathrm{TEMP}}_t^e = \frac{\beta_2 + \gamma_2}{1 - \beta_1 - \gamma_1}\Delta\overline{\mathrm{RAD}}_t + \frac{\gamma_3}{1 - \beta_1 - \gamma_1}\Delta \log(\mathrm{CO2}_t), \tag{2.9}$$

where the first term represents the change in the steady state temperature due to a change in solar radiation (for example, caused by dimming), while the second term represents the change in the steady state temperature due to a change in CO2. In a similar way, we can calculate decompositions at a local level or at a partially aggregated level (such as a continent).

Again using (2.7) and (2.8), we can rewrite Equations (2.5) and (2.6) as

$$\mathrm{TEMP}_{i,t+1} - \mathrm{TEMP}_{it} = (1-\beta_1)\left(\mathrm{TEMP}_{it}^e - \mathrm{TEMP}_{it}\right) - \gamma_1\left(\overline{\mathrm{TEMP}}_t^e - \overline{\mathrm{TEMP}}_t\right),$$

which reveals that the system is mean-reverting (as long as $\beta_1 \leqslant 1$, $\gamma_1 \leqslant 0$, and the steady state temperatures are taken as the 'means'), where $-\gamma_1$ quantifies the speed of mean reversion for deviations from the global steady state temperature, and $1 - \beta_1$ quantifies the speed at the local level.

2.4.3 Uncertainty

In a world without uncertainty, the development of temperature over time and weather stations is assumed to be determined by Equations (2.5) and (2.6), where $i = 1, \ldots, N$ indexes the weather station ($N = 1337$) and $t = 1, \ldots, T$ the year ($T = 44$). There is, however, considerable uncertainty about nonlinearities, omitted variables, and many other issues. Uncertainty is introduced through three channels. We have a station-specific effect α_i, which captures any effects specific for weather station i, not changing over time (at least, not changing over the sample period); a time-specific effect η_t, which captures those station-independent time effects not captured by $\overline{\text{TEMP}}_t$, $\overline{\text{RAD}}_t$, and $\log(\text{CO2}_t)$; and a station-specific and time-dependent idiosyncratic effect u_{it}. Introducing these three error terms results in the following econometric specification for weather station i at year t:

$$\text{TEMP}_{i,t+1} = \beta_1 \text{TEMP}_{it} + \beta_2 \text{RAD}_{it} + \alpha_i + \lambda_t + u_{i,t+1}, \tag{2.10}$$

$$\lambda_t = \gamma_0 + \gamma_1 \overline{\text{TEMP}}_t + \gamma_2 \overline{\text{RAD}}_t + \gamma_3 \log(\text{CO2}_t) + \eta_t. \tag{2.11}$$

Once the parameters in the two equations have been estimated, the steady state temperatures and the decompositions discussed in the previous subsection can be calculated straightforwardly.

In order to estimate the parameters in (2.10) and (2.11) we need to impose distributional assumptions. In our specification there is cross-sectional dependence via the time effects λ_t. To deal with this dependence, we consider (2.10) conditional on λ_t. Given λ_t, we assume independence over the weather stations. The λ_t will then capture cross-sectional correlation. We shall make distributional assumptions similar to those proposed in Arellano and Bond (1991), and Blundell and Bond (1998), and this allows us to estimate (in a

first round) the β-parameters in (2.10) and also the time effects λ_t, using standard panel data estimation techniques. Next, given the estimated time effects, we use (2.11) together with the usual linear regression assumptions to estimate the γ-parameters in a second round by ordinary least squares.

We now describe the distributional assumptions that we impose on (2.10), in addition to assuming independence over weather stations, conditional on the time effects. For each weather station i and time period t in our dataset we shall assume:

$$\mathrm{E}[\alpha_i + u_{it}] = 0, \tag{A1}$$

$$\mathrm{E}[u_{i,t-s}(\alpha_i + u_{it})] = 0 \ \ (s \geqslant 1), \tag{A2}$$

$$\mathrm{E}[\Delta \mathrm{RAD}_{i,t-s} \Delta u_{it}] = 0 \ \ (s \geqslant 1), \tag{A3}$$

$$\mathrm{E}[\mathrm{TEMP}_{i,t-s} \Delta u_{it}] = 0 \ \ (s \geqslant 2), \tag{A4}$$

$$\mathrm{E}[\Delta \mathrm{TEMP}^e_{i,t-s}(\alpha_i + u_{it})] = 0 \ \ (s \geqslant 1). \tag{A5}$$

Assumptions (A1) and (A2) are standard zero mean and zero correlation assumptions for the station-specific and idiosyncratic error terms. Assumptions (A3) and (A4) are standard zero correlation assumptions between independent or lagged dependent variables and error terms. Assumption (A5) concerns the change in steady state temperature, and states that future error terms do not deviate systematically with this change. Moreover, we assume for some $\tau \leqslant 1$, possibly far back in the past and independent of i,

$$\mathrm{TEMP}_{i,\tau} = \mathrm{TEMP}^e_{i,\tau}. \tag{A6}$$

This assumption can be seen as an initial condition, stating that the system was in a steady state at some point in the past.

2.4.4 Correlation

Even though (conditional on the time effects) the idiosyncratic errors u_{it} are assumed to be independent over weather stations and have to satisfy (A2),

the complete error term in (2.10)–(2.11) equals $\alpha_i + \eta_t + u_{i,t+1}$. This implies that cross-sectional and time correlation is built into the model, and we illustrate this fact under additional mean-independence assumptions (which imply Assumption (A1)). We first consider correlation over time, and we write $\operatorname{cov}(\mathrm{TEMP}_{i,t+1}, \mathrm{TEMP}_{it}) = C_1 + C_2$, where

$$\begin{aligned} C_1 &= \operatorname{cov}\left(\mathrm{E}\left(\mathrm{TEMP}_{i,t+1} \,|\, \mathcal{I}_{it}\right), \mathrm{E}\left(\mathrm{TEMP}_{it} \,|\, \mathcal{I}_{it}\right)\right), \\ C_2 &= \mathrm{E}\left(\operatorname{cov}\left(\mathrm{TEMP}_{i,t+1}, \mathrm{TEMP}_{it} \,|\, \mathcal{I}_{it}\right)\right) \end{aligned}$$

represent the covariance captured by the systematic part, and the covariance due to the error terms (conditional upon $\mathcal{I}_{it}$), respectively, and

$$\mathcal{I}_{it} = \left\{\mathrm{TEMP}_{i,t-1}, \mathrm{RAD}_{it}, \mathrm{RAD}_{i,t-1}, \mathrm{CO2}_t, \mathrm{CO2}_{t-1}, \overline{\mathrm{TEMP}}_{t-1}, \overline{\mathrm{RAD}}_t, \overline{\mathrm{RAD}}_{t-1}\right\}.$$

denotes the conditioning set. We are interested in C_2 and we shall show in Section 2.5.1 that C_2 is relatively small. The additional mean-independence assumption is $\mathrm{E}\left(\alpha_i + \eta_t + u_{i,t+1} \,|\, \mathcal{I}_{it}\right) = 0$, which implies that the average conditional expectation equals the unconditional expectation. Given our distributional assumptions,

$$\begin{aligned} C_2 = {} & \beta_1 \operatorname{var}\left(\alpha_i + u_{it}\right) + \gamma_1 \operatorname{cov}\left(\overline{\alpha} + \overline{u}_t, \alpha_i + u_{it}\right) + \left(\beta_1 + \gamma_1\right) \operatorname{var}\left(\eta_{t-1}\right) \\ & + \operatorname{var}\left(\alpha_i\right) + \operatorname{cov}\left(\alpha_i, u_{i,t+1}\right). \end{aligned} \tag{2.12}$$

This shows that the error structure generates time correlation in two ways, due to the autoregressive nature of the model ('state dependence') captured by the first three terms (if $\beta_1 \neq 0$ or $\gamma_1 \neq 0$), and due to the correlation of the individual effect with itself and with the idiosyncratic error term ('unobserved heterogeneity') captured by the final two terms.

Next, we consider spatial correlation. We decompose

$$\operatorname{cov}(\mathrm{TEMP}_{i,t+1}, \mathrm{TEMP}_{j,t+1})$$

in the same way as before, but with a different conditioning set, namely

$$\widetilde{\mathcal{I}}_{ijt} = \left\{ \mathrm{TEMP}_{it}, \mathrm{RAD}_{it}, \mathrm{TEMP}_{jt}, \mathrm{RAD}_{jt}, \overline{\mathrm{TEMP}}_t, \overline{\mathrm{RAD}}_t, \mathrm{CO2}_t \right\}.$$

The mean-independence assumption now reads $\mathrm{E}(\alpha_i + \eta_t + u_{i,t+1} \mid \widetilde{\mathcal{I}}_t) = 0$, and, using our distributional assumptions, the second term in the covariance decomposition is equal to $\mathrm{var}\,(\eta_t)$. Thus, the error term in the time effect captures the error-term-specific cross-sectional correlation.

2.4.5 Moment restrictions with missing observations

Some solar radiation observations are missing and this may cause a selection problem. We now describe how the distributional assumptions (A1)–(A6) can be manipulated to construct moment restrictions such that the parameters in (2.10) can be estimated by the Generalized Method of Moments (GMM) in the presence of missing observations.

We introduce selection variables r_{it}, such that $r_{it} = 0$ if observation (i, t) on solar radiation is missing, and $r_{it} = 1$ if the observation is present. Conditional on the time effect, we combine the distributional assumptions (A1)–(A6) with the assumption that the missing observations are MCAR, except possibly for the level. By this we mean that, under the assumption that the selection variables are independent of the random variables appearing in (A1)–(A6), the moment restrictions are valid in terms of the parameters appearing in (2.10), except possibly for the level. Since the level will be captured by the time effects λ_t, our assumption implies that we may not be able to estimate the level of the time effects consistently, but we will be able to estimate, for example, $\lambda_t - \lambda_1$ consistently.

We use the following moment restrictions in estimating the parameters of (2.10):

$$\mathrm{E}\sum_{t=2}^{T}\left[r_{i,t-1}(\alpha_i + u_{it})\right] = 0, \tag{M1}$$

$$\mathrm{E}\big[r_{i,t-1}r_{i,t-2}\Delta u_{it}\big] = 0 \;\; (t = 3, \ldots, T), \tag{M2}$$

$$\mathrm{E}\sum_{t=3}^{T}\left[r_{i,t-1}r_{i,t-2}\Delta\mathrm{RAD}_{i,t-1}\Delta u_{it}\right] = 0, \tag{M3}$$

$$\mathrm{E}\big[r_{i,t-1}r_{i,t-2}\mathrm{TEMP}_{i,t-s}\Delta u_{it}\big] = 0 \;\; (t = 3, \ldots, T;\; s = 2, \ldots, \min(t-1, 4)), \tag{M4}$$

$$\mathrm{E}\big[r_{i,t-1}(\alpha_i + u_{it})\Delta\mathrm{TEMP}_{i,t-1}\big] = 0 \;\; (t = 3, \ldots, T). \tag{M5}$$

Restrictions (M1) and (M2) are derived from (A1) and the MCAR assumption, where (M2) is obtained by taking time differences of (A1). Restrictions (M3) and (M4) are derived from (A3) and (A4), respectively, together with the MCAR assumption. Restriction (M5) follows from taking time differences of (2.10) (until reaching $t = \tau$), combined with (A2), (A3), the initial condition (A6), and the MCAR assumption. The restrictions (M1)–(M4) are based on the moment conditions in Arellano and Bond (1991); the additional restriction (M5) is based on Blundell and Bond (1998).

The first round provides consistent estimates of $\lambda_t - \lambda_1$ $(t = 2, \ldots, T-1)$, and we use these estimates in Equation (2.11). We calculate the global averages of both temperature and solar radiation, using the differences in the unbalanced panel in the following way. Let $\overline{\mathrm{TEMP}}_1$ be the global average temperature in the first year of the 'complete panel' (the panel including the missing observations), and let $\overline{\mathrm{RAD}}_1$ be the global average solar radiation in the first year of the 'unbalanced panel' (the panel without the missing observations). Then, $\overline{\mathrm{TEMP}}_t$ is calculated as

$$\overline{\mathrm{TEMP}}_t = \overline{\mathrm{TEMP}}_{t-1} + \frac{1}{\sum_{i=1}^{N} r_{it}r_{i,t-1}}\sum_{i=1}^{N} r_{it}r_{i,t-1}\Delta\mathrm{TEMP}_{it}, \tag{2.13}$$

for $t = 2, \ldots, T$. $\overline{\mathrm{RAD}}_t$ is calculated similarly.

When estimating (2.11) we impose the usual linear regression assumptions, and we assume that applying least squares yields unbiased estimates, except again for the level. This implies that the constant term may be biased. When calculating the standard errors of the linear regression coefficients, we ignore the first-round inaccuracy, because the number of observations in the first round (N weather stations) is much larger than the number of observations in the second round ($T - 1$ years).

2.5 Empirical results

We now present the empirical results. In Section 2.5.1 we discuss the estimation results. In Section 2.5.2 we investigate the 1991 eruption of Mount Pinatubo to test the performance of our model. In Sections 2.5.3 and 2.5.4 we present the decomposition of the temperature change into a greenhouse and a solar radiation effect, both in terms of observed and steady state temperatures. We also consider this decomposition at regional levels (continents).

2.5.1 Parameter estimates

The estimation results for our model, based on Equations (2.10) and (2.11), are presented in Table 2.3. The first two columns give the estimates and standard errors of the β's in Equation (2.10), while the next three columns contain the estimates and standard errors of the γ's in Equation (2.11). All estimates have the expected signs and are statistically significantly different from zero (at the 5% level). The panel-data based estimates of Equation (2.10) are far more accurate than the time-series based estimates of Equation (2.11), and this supports our approach to ignore the first-round inaccuracy in the second round. In the subsequent subsections we shall use these parameter estimates to characterize our climate model.

For a dynamic model such as our econometric model, it is standard practice to use the Arellano-Bond estimator, that is, to apply GMM to the moment

TEMP_{it} (β_1)	RAD_{it} (β_2)	$\overline{\text{TEMP}}_t$ (γ_1)	$\overline{\text{RAD}}_t$ (γ_2)	$\log \text{CO2}_t$ (γ_3)
0.9063	0.0087	−0.8235	0.0614	10.6955
(0.0046)	(0.0008)	(0.1839)	(0.0219)	(2.3958)

Table 2.3: Parameter estimates and standard errors

restrictions (M1)–(M4); see Arellano and Bond (1991). This estimator performs poorly, however, when the autoregressive coefficient β_1 or the variance ratio $\text{var}(\alpha_i)/\,\text{var}(u_{it})$ is large (Blundell and Bond, 1998). Including moment restriction (M5) may then yield better results. In our case the estimate of the autoregressive coefficient is $\hat{\beta}_1 = 0.91$ and the estimate of the variance ratio is 0.98. Both are 'large', thus motivating our choice to use all moment restrictions (M1)–(M5).

In terms of the implied correlation structure as described in Section 2.4.4, we estimate that the temporal correlation, calculated from (2.12), is 0.017 with 0.011 due to state dependence and 0.006 to unobserved heterogeneity. Since the total temporal correlation is 0.996, the error terms contribute only a small part; most is captured by the systematic part of the model. The estimate of the final term in (2.12), $\text{cov}\,(\alpha_i, u_{i,t+1})$, is very close to zero, implying that, given the assumptions in Section 2.4.4, the autocorrelation in the idiosyncratic error terms u_{it} is also estimated to be zero (using (A2)). The cross-sectional correlation, given by $\text{var}\,(\eta_t)\,/\,\text{var}\,(\text{TEMP}_{i,t+1})$, is estimated to be 0.002, and the estimate of the total cross-sectional correlation is 0.16. Again, the contribution of the error terms is small.

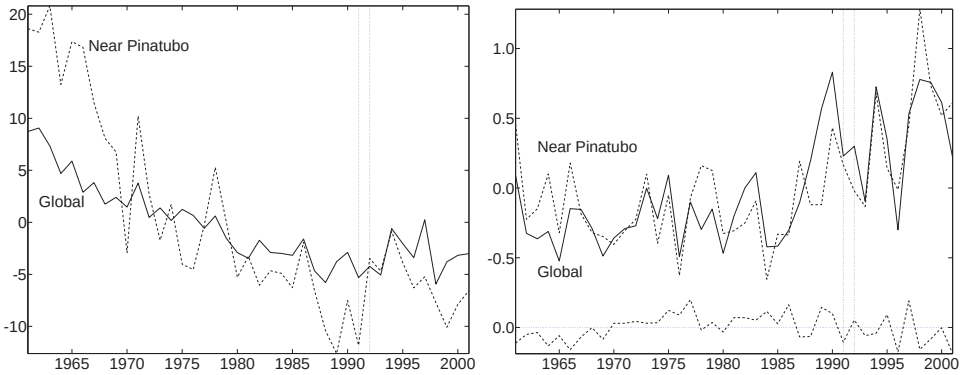
Using the estimated β's and γ's we can investigate whether dimming is a local or a global effect or both. If $H_0 : a_1 = a_2$ holds then dimming is only a local effect. In terms of our reduced-form parameters we need to test $H_0 : \gamma_2 = 0$. Since $\hat{\gamma}_2$ is significantly different from zero, we reject H_0 and conclude that there is evidence for a global dimming effect. On the other hand, if $H_0 : a_2 = 0$ holds then dimming is *only* global. Here we need to test $H_0 : \beta_2 = 0$ and this is also rejected. Hence, we find both a local and a global dimming effect, but since a_1 is much larger than a_2, the local effect is much more important than the global effect.

The specification (2.10)–(2.11) is linear in the independent variables. This linear specification should be seen as a linear approximation to a nonlinear structure. To test the validity of the linear approximation, we performed a number of specification tests. In particular, we calculated the in-sample predictions according to the specification (2.10)–(2.11), and compared these to three in-sample predictions, where in each case one of the linear terms in (2.11) was replaced by a fully flexible specification in this variable, estimated nonparametrically using Robinson's (1988) semiparametric regression approach. Only in case of CO2 do we find some statistically significant differences between our linear specification and the alternative partial nonparametric regression in-sample predictions, indicating that, at least in-sample, the linear specification performs well.

2.5.2 Mount Pinatubo

How confident can we be that our results are driven by and identified in the data, and not just an artifact of model choice? A natural environment for studying this question is to consider a shock in one of the explanatory variables, say solar radiation. If the model is correctly specified, then this should lead to a shock in the prediction of the dependent variable (temperature), but not to a shock in the residuals. A large volcanic eruption provides the ideal environment, and the June 1991 eruption of Mount Pinatubo on the island of Luzon in the Philippines was the largest eruption in our data period, in fact the largest disturbance of the stratosphere since the eruption of Krakatau in 1883. An estimated 30 Teragrams (Megatonnes) of aerosols were released into the atmosphere.

Figure 2.4 summarizes our analysis. In panel (a) we present the solar radiation time series for the 100 stations closest to Mount Pinatubo ('Near Pinatubo') and compare this series with the solar radiation time series for all stations in our dataset ('Global'). Both series are normalized so that their average over the period is zero. The two vertical lines indicate the years 1991 (the year of the eruption) and 1992. The 'Pinatubo effect' is clearly visible:



(a) Solar radiation **(b)** Temperature and residuals

Figure 2.4: Analysis of the Mount Pinatubo eruption

the global average in 1991 is 5.33 Wm^{-2} lower than the average over 1959–1990, and near the Pinatubo even 12.95 Wm^{-2} lower. This effect is largest near Mount Pinatubo, since the eruption lasted until August, with episodic eruptions in September. But there is also a global effect due to the fast dispersion of the aerosols across the globe: the aerosol cloud moved westward and circled the globe in approximately 22 days (McCormick et al, 1995).

Our model predicts that there should be a temperature shock in 1992, and this negative effect on temperature is visible from panel (b), not just in 1992 but also in 1993. We should be a little careful in our conclusions, because both solar radiation and temperature are volatile (especially the graphs based on only 100 stations).

The key graph is at the bottom of panel (b) where we plot the (scaled) residuals, averaged over the stations close to Mount Pinatubo. There is no sign of any anomaly in the residuals. It seems justified therefore to have confidence that our results are driven by and identified in the data.

2.5.3 Greenhouse and solar radiation effects

The purpose of this paper is to try and decompose the observed (in-sample) total change in temperature into a change that can be attributed to a change in the concentration of greenhouse gases, and a change caused by a change in the solar radiation reaching the surface. Our econometric model enables us to do

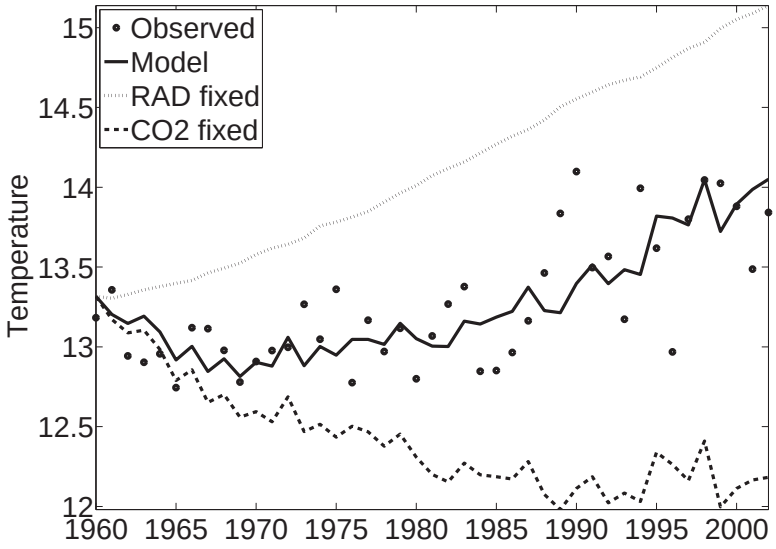


Figure 2.5: Decomposition of temperature change, 1960–2002

this, and Figure 2.5 illustrates the resulting decomposition. The dots represent the observed global average temperature, calculated using Equation (2.13), and setting $\overline{\text{TEMP}}_1$ equal to the average temperature in the first year of the complete panel. The solid curve gives the expected global average temperature according to our model, conditional on the observed development of carbon dioxide and solar radiation. We set the level of this curve such that its time average equals the time average of the observed temperature series. The in-sample change in average temperature equals 0.66 °C (1960–2002), while the model predicts the slightly higher temperature change of 0.73 °C. The solid curve follows the actual series closely, and hence our model is able to reproduce the pattern of in-sample temperature changes well.

Two further temperature series are presented in Figure 2.5, and these represent the decomposition. The lower curve shows the expected temperature if carbon dioxide is assumed to remain at its 1959 level (the start of our dataset). The upper curve shows the expected temperature if solar radiation is assumed to remain at the level of 1959. The difference between the lower curve and the solid curve can be interpreted as the greenhouse effect for the period 1959–2002, while the difference between the upper curve and the solid curve can be interpreted as the solar radiation effect. The figure shows that, without

the increase in greenhouse gases, the expected global average temperature would have been 1.87 °C lower (with standard error 0.32): the greenhouse effect. Also, if global average solar radiation is unchanged from its initial level, then the expected global average temperature would have been 1.09 °C higher (standard error 0.31): the solar radiation effect. The predicted temperature change of 0.73 °C thus decomposes as $0.73 = 1.87 - 1.09 - 0.05$, where 0.05 is a remainder term due to the fact that we are not in a steady state. We conclude that the solar radiation effect is important, masking 58% of the increase due to the greenhouse effect.

Let us compare these findings with the literature. Such a comparison should be interpreted with some care, because existing studies use different time periods than our study, and some focus on specific regions. Furthermore, our solar radiation effect includes factors other than aerosols that influence the amount of incoming solar radiation. Taking these caveats into account, we find that the existing findings broadly agree with ours. Tett et al (2002) report a greenhouse effect of 0.9 °C per century. Stott et al (2006) find that 0.7–1.3 °C of warming is due to greenhouse gases, and that 0.33–0.49 °C of cooling is due to aerosols. Allen et al (2006) find that the twentieth century greenhouse effect is in the range of 0.3–1.2 °C, with a cooling of 0.7 °C due to aerosols. Our results imply a more important greenhouse effect.

Regarding the solar radiation masking effect, Crutzen and Ramanathan (2003) report a masking effect of 45% from 1850 to the present. Applying their reasoning to the results in Anderson et al (2003) yields values in the range 37%–56% for the same time period. Similarly, applying their reasoning to Bellouin et al (2005) and Myhre (2009) yields values of 70% and 11%, respectively. For 1930–2002, Ramanathan et al (2005) find that aerosols may have masked as much as 50% of the surface warming due to the global increase in greenhouse gases. Our findings in terms of the relative importance of the solar radiation effect are in line with this literature.

Actual changes may be different from steady state changes to which they will converge. Therefore we investigate the steady state effects next.

2.5.4 Steady state effects

We decompose the steady state temperature change in the period 1960–2002 into a solar radiation and a greenhouse effect, both globally and regionally, at the level of continents. At the global level, the change in average steady state temperature equals 0.92 °C (standard error 0.18). The global average steady state temperature would have been 1.90 °C (0.35) lower without the increase in CO2, while the average steady state temperature would have been 0.98 °C (0.31) higher if global average solar radiation would still be at its initial level. Notice that the decomposition in steady state contains no remainder term: $0.92 = 1.90 - 0.98$. Our results imply that the global mean-reverting coefficient $-\gamma_1$ equals 0.82 (0.18). The mean-reverting speed at the global level is therefore high, and convergence to the global steady state temperature is fast.

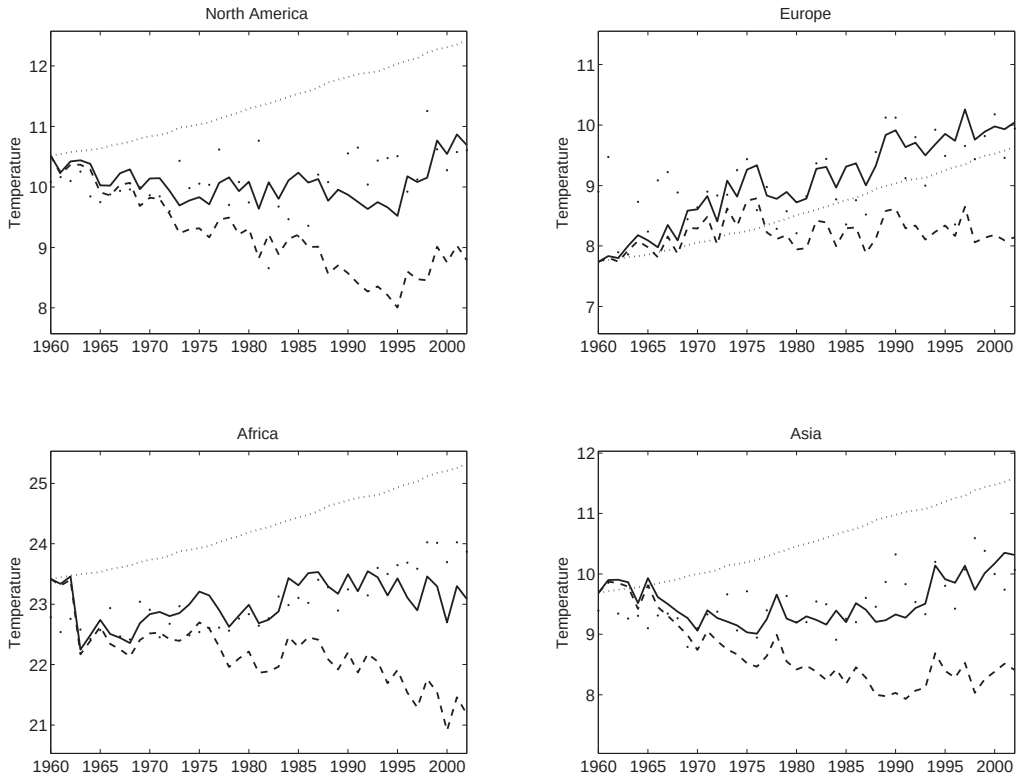


Figure 2.6: Decomposition of temperature change by continent, 1960–2002

At the regional (continent) level, the changes in steady state temperature may differ, due to local dimming. These regional effects, calculated using (2.8), are illustrated in Figure 2.6, where we show the decomposition for four continents: Africa, Asia, Europe, and North America. The graphs are similar to Figure 2.5, except that the curves now show steady state temperatures. In North America the average steady state temperature would have been 1.73 °C higher in the case where solar radiation would still be at the 1960 level. In Asia the temperature would be 1.73 °C higher, and in Africa even 2.23 °C. The uncertainty of these effect estimates are similar to those in Figure 2.5, since they are based on the same parameter estimates. One would perhaps expect that the solar radiation effect in Asia becomes larger in comparison with North America in the 1990s, due to the expansion of the Asian economies and the associated increase in sulfur emissions. However, external data on sulfur emissions reveal that Chinese sulfur emissions leveled off after 1989, and this is consistent with Figure 2.6. These results demonstrate that the local solar radiation effect may be different from the global effect, and also much more important than the greenhouse effect, masking even more than 100% of the temperature increase due to the greenhouse effect. The local mean-reverting coefficient $1-\beta_1$ equals 0.094 (standard error 0.005). The local mean-reverting speed is thus much lower than the global mean-reverting speed, implying that convergence at local levels can be slow.

2.6 Sensitivity analysis

Our benchmark model is based on a large number of assumptions, in particular about the climate model, about the statistical model, and about the data. Any or all of these assumptions may be incorrect. In this section we ask whether small deviations from our assumptions will cause large or small changes in our conclusions. In the former case the conclusions are apparently sensitive to a particular assumption; in the latter case they are not. Obviously we prefer that our conclusions are not sensitive, but this is something that needs to be investigated, especially in the context of climate change where there is much

uncertainty about the process. We organize our sensitivity analyses in three groups: climate model issues, statistical model issues, and data issues. In our sensitivity analysis we focus on Figure 2.5, that is, we ask the following question: How sensitive to our assumptions is the decomposition of the total temperature change into a change due to greenhouse gases represented by CO2 (the greenhouse effect) and a change due to dimming (the solar radiation effect)? Table 2.4 summarizes our results.

2.6.1 Climate model issues

We consider two ways to change the climate model. The first is to make the solar radiation effect latitude-dependent. The second is to consider a static model.

	Method		Solar radiation	Greenhouse
1	Benchmark		−1.09 (0.31)	1.87 (0.32)
Climate model issues				
2a	Albedo		−0.92 (0.34)	2.34 (0.41)
2b			−1.20 (0.29)	2.24 (0.28)
3	Static		−0.78 (0.15)	1.59 (0.17)
Statistical model issues				
4a	Lags	Two lags	−1.05 (0.31)	1.84 (0.32)
4b		Four lags	−1.08 (0.31)	1.88 (0.32)
5	Arellano-Bond		−0.78 (0.29)	1.73 (0.30)
6	One round		−0.07 (0.03)	1.08 (0.03)
Data issues				
7	Definition of $\overline{\text{TEMP}}$		−1.16 (0.25)	1.78 (0.24)
8	Spatial Independence		−1.07 (0.32)	1.95 (0.33)
9	Weights		−1.43 (0.28)	1.71 (0.25)
10a	1/2 most complete stations		−0.88 (0.30)	1.73 (0.30)
10b	2/3 most complete stations		−1.10 (0.31)	1.86 (0.31)

Table 2.4: Sensitivity analysis: solar radiation and greenhouse effects

In our benchmark model we made the assumption that the solar radiation effect is the same for each weather station. One might argue however that the solar radiation effect depends on the latitude, due to a latitude-specific albedo effect. We investigate two methods to allow for this dependency.

In the first method (model 2a), we divide the Earth into six latitude zones of equal size. We let $\text{RAD}_{it}^{l} = \text{RAD}_{it}$ if station i is in zone l, and 0 otherwise ($l = 1, \ldots, 6$), and we replace $\beta_2 \text{RAD}_{it}$ in (2.10) by $\sum_{l=1}^{6} \beta_{2l} \text{RAD}_{it}^{l}$. We find that all radiation coefficients are positive, and that they are lower for zones further away from the equator. The implications for the decomposition are that, compared to our benchmark results, the solar radiation effect decreases and the greenhouse effect increases.

In the second method (model 2b), we let the radiation coefficient be a linear function of the distance to the equator, that is, $\beta_{2,i} = a_0 + a_1 |\text{LAT}_i/90|$, where a_1 is allowed to be different per hemisphere. We find that both the solar radiation effect and the greenhouse effect increase. Hence, if we assume that the solar radiation effect is latitude-dependent, then the magnitude of the solar radiation effect does not change systematically, but may become smaller or larger than the benchmark, depending on the way the dependence on latitude is modeled. But since in both models the greenhouse effect increases, we find that the solar radiation effect only masks 39% or 53% of the increase due to the greenhouse effect.

Our climate model is based on the idea that a surplus or a deficit in the energy balance causes a change in temperature. This results in our dynamic specification (2.10)–(2.11). Alternatively, one could set up a climate model by linking the temperature to the energy level. Such an approach leads to a static panel data model, for example our model (2.10)–(2.11), but then with $\beta_1 = \gamma_1 = 0$ and with TEMP_{it} as dependent variable instead of $\text{TEMP}_{i,t+1}$. We estimate this static model (model 3) imposing moment restrictions analogous to the benchmark model. We find lower solar radiation and greenhouse effects, where the solar radiation effect becomes, relatively speaking, somewhat less important (49%). Without a dynamic autoregressive part, the individual station-specific effect becomes much more important than in the benchmark model, capturing 0.918 (instead of 0.012) of the total temporal autocorrelation of 0.996. In this case the individual effects also capture some of the station-specific trends over time, leading to lower solar radiation and greenhouse effects. Overall, we conclude that the decomposition of the total temperature change into a change due to greenhouse gases represented by

CO2 (the greenhouse effect) and a change due to dimming (the solar radiation effect) is not very sensitive to our assumptions.

2.6.2 Statistical model issues

We investigate the sensitivity of the decomposition with respect to three deviations in the statistical model. First, for restriction (M4), we have chosen a maximum of three lags of TEMP to be used as instruments. We consider as alternatives two lags (model 4a) and four lags (model 4b). This has only a small effect on the decomposition results. Second, we use the moment restrictions (M1)–(M4) in our benchmark model, based on Arellano and Bond (1991), extended with the moment restriction (M5) as in Blundell and Bond (1998). Model 5 is obtained by estimating the model using only (M1)–(M4). Even though the underlying parameter estimates change significantly, the results in terms of the decomposition are close to those of the benchmark model.

Third, we consider a restricted version of our benchmark model, where we do not estimate the model in two rounds, but in one round (model 6). We use Equations (2.10)–(2.11), but set the time-specific parameter to zero, thus ignoring possible cross-sectional correlations. We estimate the model using the moment conditions (M1)–(M5). In terms of the decomposition, we find a substantial decrease in the greenhouse effect, while the solar radiation effect becomes quite small (although still statistically significantly different from zero). The high accuracy of the estimates is due to the single-round estimation, based solely on the large number of weather stations. Without the time-specific intercepts, the imposed time structure does not seem to allow for sufficient flexibility, resulting in findings quite different from the other specifications.

2.6.3 Data issues

Finally, we consider four data issues. In the benchmark model we have calculated the mean temperatures $\overline{\text{TEMP}}_t$ and the mean solar radiation levels $\overline{\text{RAD}}_t$ using differences in the unbalanced panel, in order to avoid potential

sample selection problems caused by missing observations. But these averages can be calculated in various ways. In model 7 we take, as an alternative, the following temperature and solar radiation means in the second round:

$$\overline{\text{TEMP}}_t = \frac{1}{N}\sum_{i=1}^{N} \text{TEMP}_{it}, \qquad \overline{\text{RAD}}_t = \frac{\sum_{i=1}^{N} r_{i,t+1}\text{RAD}_{it}}{\sum_{i=1}^{N} r_{i,t+1}}.$$

Thus we take the average in year t in the complete panel to calculate $\overline{\text{TEMP}}_t$, and the average in year t in the unbalanced panel to calculate $\overline{\text{RAD}}_t$. This changes the levels, in particular the level of temperature. The corresponding decomposition effects (which are changes) are close to the benchmark. Hence, the alternative way of calculating the means affects the levels, but not the changes in a statistically significant way, and this is in line with our assumption that the unbalanced sample is representative for the complete panel in terms of (temperature) changes.

When we calculate the spatial correlation using the model-based idiosyncratic error terms $u_{i,t+1}$ and $u_{j,t+1}$, we find that this correlation is negligible for weather stations further apart, in line with our assumptions. Only for weather stations close to each other, we find spatial correlation, which disappears rapidly with increasing distance. This spatial correlation between weather stations that are close is due to the construction of the dataset, where weather stations in the same grid cell share the same temperature data. To see whether our decomposition results are sensitive to this spatial correlation in the idiosyncratic error terms of nearby weather stations, we consider a subsample of our sample, by drawing randomly one weather station from each temperature grid cell. This reduces the number of weather stations by 153, while the number of observations becomes 16949 instead of 18395 (model 8). The resulting changes in the solar radiation and greenhouse effects are minor.

In the benchmark model we assume a random sample, conditional upon the time effects. However, the weather stations are not evenly spread over the continents. For example, the ratio of South American weather stations to its landmass is too low, while for Europe it is too high. To deal with this uneven spread of weather stations over the continents, we estimate a weighted version

(model 9) of the benchmark model, with weights w_i $(i = 1, \ldots, N)$ defined as the proportional size divided by the proportional number of observations of the continent where station i is located. We adapt the definition of $\overline{\text{TEMP}}_t$ and $\overline{\text{RAD}}_t$ accordingly. In this model, the solar radiation effect is larger (and estimated more accurately), while the greenhouse effect is slightly smaller (and also estimated more accurately). However, we find no statistically significant differences between the decomposition effects of the weighted and unweighed versions.

For most weather stations we do not have full records on solar radiation during the whole sample period. For some weather stations we observe solar radiation only during some years, while for other weather stations we observe solar radiation during most years. Our assumption is that this unbalanced structure of our panel is not causing a selection effect. A recommended way to check this, is to compare the estimation results with a more balanced subpanel, including only the weather stations with (more) complete records; see Verbeek and Nijman (1992). We consider the more balanced subpanel, containing one-half of the weather stations with the most complete solar radiation records (model 10a). Both the solar radiation effect and the greenhouse effect become smaller. As a result, the solar radiation effect now masks 51% (instead of 58% in the benchmark model) of the increase due to the greenhouse effect. If we chose 2/3 instead of 1/2, then the results in Table 2.4 (model 10b) are almost identical to our benchmark results. The missing observations do therefore have an effect on our results, as one would expect, but this effect is small.

2.7 Conclusions

In this paper we propose a climate model based on the Earth's energy balance. We then modify this climate model to obtain an econometric model, and we estimate its parameters using dynamic panel data methods. Our data consist of solar radiation, temperature, and carbon dioxide concentrations from 1337 weather stations around the world for the period 1959–2002.

During the 43 years 1960–2002 temperature increased by an estimated 0.73 °C, which we decompose as $0.73 = 1.87 - 1.09 - 0.05$, namely a greenhouse effect of 1.87 °C (standard error 0.32), a solar radiation effect of 1.09 °C (0.31), and a remainder term of 0.05. Hence, if aerosols and solar radiation would have remained at the 1959 level, then the expected global average temperature would have been 1.09 °C higher. The solar radiation effect is therefore important, masking 58% of the increase due to the greenhouse effect. Ignoring dimming thus causes a serious underestimation of the greenhouse effect.

Our approach has several strengths and several weaknesses. The weak points are that some important climate processes (for example, carbon storage in the ocean) are not modeled; that only land stations and no sea stations are considered; and finally that data availability limits our time horizon. Some would also criticize our frequentist (as opposed to Bayesian) approach. While modeling environmental data based on Bayesian hierarchical models has become popular and such models provide a clear framework for dealing with the various aspects of the climate system and with data issues, we have not chosen for this approach because of the much more restrictive distributional assumptions that have to be made on the sources of uncertainty, and on the variable that contains the missings.

The strong points are that our model is simple enough to allow estimation rather than calibration of the reduced-form parameters and their uncertainties, that the reduced-form parameters are all that is needed for our analysis, and that analysis at all levels of aggregation is possible. Our main result is contained in Figure 2.5, where we present the decomposition in greenhouse and solar radiation effects. An important aspect of the paper is the sensitivity analysis. We present not only Figure 2.5, but we also ask how the figure would change if we make small adjustments to our underlying assumptions. Climate models are often criticized for not being robust. Extensive sensitivity analysis demonstrates that our conclusions are relatively robust against small changes in a variety of assumptions.

Chapter 3

Expected utility and catastrophic risk in a stochastic economy-climate model

Abstract: We specify a stochastic economy-climate model using expected power utility and explicitly demonstrate its fragility to heavy-tailed distributional assumptions. We derive necessary and sufficient conditions on the utility function to avoid fragility and solve our stochastic economy-climate model for two examples of compatible utility functions. We further develop and implement a procedure to learn the input parameters of our model and show that the model thus specified produces quite robust optimal policies. The numerical results indicate that higher levels of uncertainty lead to less abatement and consumption, and to more investment, but this effect is not unlimited.

3.1 Introduction

An economist, when asked to model decision making under risk and uncertainty for normative purposes, would typically work within the expected utility framework with constant relative risk aversion (that is, power utility). A statistician, on the other hand, would model economic catastrophes through probability distributions with heavy tails. Unfortunately, expected utility is fragile with respect to heavy-tailed distributional assumptions: expected utility may fail to exist or it may imply conclusions that are 'incredible'.

Economists have long been aware of this tension between the expected utility paradigm and distributional assumptions (Menger, 1934), and the discussions in Arrow (1974), Ryan (1974), and Fishburn (1976) deal explicitly with the trade-off between the richness of the class of utility functions and the generality of the permitted distributional assumptions. Compelling examples in Geweke (2001) corroborate the fragility of the existence of expected power utility with respect to minor changes in distributional assumptions.

The combination of heavy-tailed distributions and the power utility family may not only imply infinite expected utility, but also infinite expected *marginal* utility, and hence, via the intertemporal marginal rate of substitution (the pricing kernel), lead to unacceptable conclusions in cost-benefit analyses. For example, with heavy-tailed log-consumption and power utility, the representative agent should postpone *any* unit of current consumption to mitigate future catastrophes. The latter aspect was recently emphasized by Weitzman (2009) in the context of catastrophic climate change. Weitzman also argues that attempts to avoid this unacceptable conclusion will necessarily be non-robust.

In this paper we study the fundamental question of how to conduct expected utility analysis in the presence of catastrophic risks, in the context of extreme climate change. Our paper is built on four beliefs, which will recur in our analysis:

Catastrophic risks are important. To study risks that can lead to catastrophe is important in many areas, for example financial (trader, insurer, bank)

distress, traffic accidents (bridge collapse, airplane crash, flight control system failure), dike bursts, killer asteroids, nuclear power plant disasters, and extreme climate change. Such low-probability high-impact events should not be ignored in cost-benefit analyses for policy making. In the context of extreme climate change: catastrophic climate changes, unlikely as they may be, should be accounted for in expected-welfare calculations for policy making.

A good model 'in the center' is not necessarily good 'at the edges'. Suppose we have estimated a function $C = a + bY$, relating consumption to disposable income. The dots in Figure 3.1 represent the data and the line gives the resulting OLS prediction $\hat{C} = \hat{a} + \hat{b}Y$. For incomes in the center, roughly be-

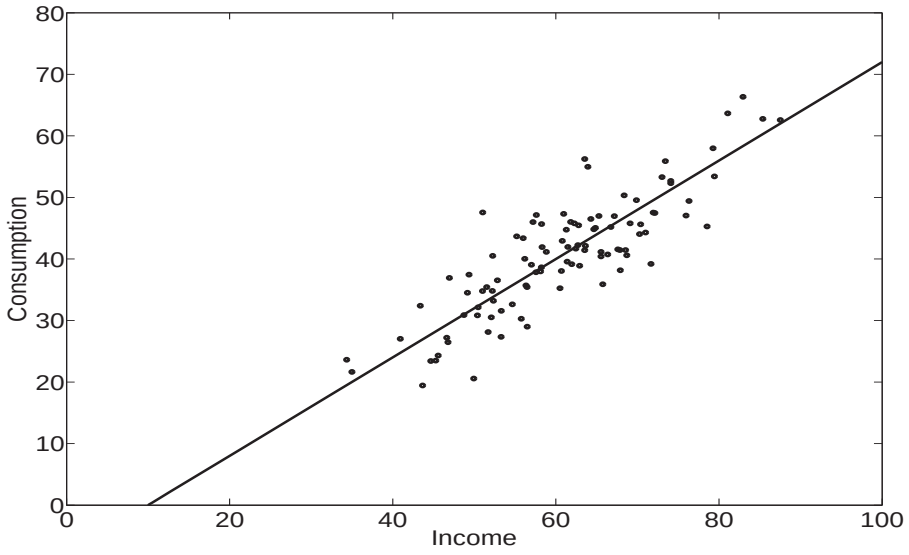


Figure 3.1: A consumption function

tween 40 and 80, the consumption function can be well approximated by the regression line. How useful is this result for very low (or very high) incomes? Not very useful. For very low incomes, predicted consumption would be negative! This does not mean that a linear consumption function is useless. But it is only useful in the center of the domain. This is simply because models are not truths but approximations, and approximations may not work well if we move too far away from the point of approximation. In our context, the widely adopted class of power utility functions, often appropriate when one considers large inputs remote from zero as is common in macroeconomics and

finance, may not work well for decision making under heavy-tailed risks with non-negligible support beyond the usual domains. Moreover, estimates of the coefficient of risk aversion are very sensitive to the particular domain of inputs that a utility function operates on (Rabin, 2000, footnote 10).

The price to reduce catastrophic risk is finite. Are we willing to spend everything to avoid children being killed at a dangerous street? Or to avoid the dikes to burst? Or a power plant to explode? Or a killer asteroid to hit the Earth? Or climate to change rapidly? No, we are not. To assume the opposite (that a society would be willing to offer all its current wealth to avoid or mitigate catastrophic risks) is not credible, not even from a normative (prescriptive, rational) perspective. In our context, there is a limit to the amount of current consumption that the representative agent is willing to give up in order to obtain one additional *certain* unit of future consumption, no matter how extreme and irreversible climate change may be. In other words: the expected pricing kernel is finite.

Light-tailed risks may result in heavy-tailed risk. When x is normally distributed (light tails) then $1/x$ has no moments (heavy tails). Also, when x is normally distributed then e^x has finite moments, but when x follows a Student distribution then e^x has no moments. In the context of extreme climate change: temperature has fluctuations but one would not expect heavy tails in its distribution. This does not, however, imply that functions of temperature cannot have heavy tails. For example, it may well be reasonable to use heavy-tailed distributional assumptions to model future (log) consumption.

There is an important literature on stochastic economy-climate models (see, for example, Keller *et al.*, 2004, Mastrandrea and Schneider, 2004, and the references therein). However, the integrated assessment models of climate economics are predominantly deterministic and rarely incorporate catastrophic risk (Ackerman *et al.*, 2010). To allow for uncertainty and extreme climate change, we start by specifying a stochastic economy-climate model that builds on Nordhaus' (2008) deterministic dynamic integrated climate and economy (DICE) model. We solve the model first with power utility and light-tailed distributional assumptions and prove that the assumption

of expected power utility is incompatible with heavy-tailed distributional assumptions. We then address the question of how to conduct expected utility analysis in the presence of catastrophic risks. In Appendix 3.C we provide necessary and sufficient conditions on the utility function, so that expected utility and expected marginal utility (hence the pricing kernel) are finite, also under heavy-tailed distributional assumptions. Restricting attention to utility functions that satisfy these compatibility conditions, we propose the two-parameter 'Burr' function as a particularly appealing utility function in our setting. We solve our stochastic economy-climate model with Burr utility (and also with the well-known exponential utility) under both light-tailed and heavy-tailed distributional assumptions.

Completing the resulting model requires specifying a number of model parameters as inputs. These parameters cannot 'simply' be determined by conventional statistical inference based on historical data. One reason is that temperature and other variables will be affected by economic policy decisions; another that economic parameters should be set so as to reflect rational decision-making behavior under circumstances that have never been encountered before. We discuss how to set the model parameters in a process towards agreement, using experts' priors on parameter values, and learning about parameters from resulting optimal model output. The key to the learning and agreement process is the translation of model parameters that are relatively difficult to interpret into quantities that allow a more straightforward interpretation. We find that our optimal policies are quite robust with respect to minor (and reasonable) changes to the input parameters.

Our numerical analysis indicates that allowing for heavy-tailed distributional assumptions in extreme climate change modeling leads to a reduction of current abatement and consumption and to an increase in current investment, when compared to a deterministic analysis. The increase in current investment may be interpreted via precautionary savings. Most notably and contrary to Weitzman (2009), while the differences are visible, they are not unlimited.

The paper is organized as follows. In Section 3.2 we propose a simplified version of Nordhaus' economy-climate model. There are two new features to

this model: scrap value functions and, more importantly, uncertainty. In Section 3.3 we specialize this model to two periods only, and maximize expected welfare with non-linear scrap value functions. We also prove that expected welfare exists under normality but not under a Student distribution (Appendix 3.B). Motivated by the fact that the expectation of the pricing kernel is finite for all outcome distributions whenever the concavity index (index of absolute risk aversion) ARA(x) is bounded (which we formally prove in Appendix 3.C), we discuss such utility functions in Section 3.4: the well-known exponential function and the less well-known 'Burr' function. Section 3.5 discusses how we can learn the parameters of our model and calibrate policy using information such as the probability of catastrophe, and reports on robustness tests. Section 3.6 concludes. There are three appendices. Appendix 3.A provides the Kuhn-Tucker conditions; Appendix 3.B contains the proof of Proposition 3.1; and Appendix 3.C discusses expected utility and tail uncertainty in a more general setting.

3.2 A simple stochastic economy-climate model

Our framework is a simple economy-climate model in the spirit of Nordhaus and Yang (1996) and Nordhaus (2008).

3.2.1 Emissions, temperature, and the economy

Everybody works. In period t, the labor force L_t together with the capital stock K_t generate GDP Y_t through a Cobb-Douglas production function

$$Y_t = A_t K_t^{\gamma} L_t^{1-\gamma} \qquad (0 < \gamma < 1),$$

where A_t represents technological efficiency and γ is the elasticity of capital. Capital is accumulated through

$$K_{t+1} = (1-\delta)K_t + I_t \qquad (0 < \delta < 1),$$

where I_t denotes investment and δ is the depreciation rate of capital. Production generates carbon dioxide (CO2) emissions E_t:

$$E_t = \sigma_t(1 - \mu_t)Y_t,$$

where σ_t denotes the emissions-to-output ratio for CO2, and μ_t is the abatement fraction for CO2. The associated CO2 concentration M_t accumulates through

$$M_{t+1} = (1 - \phi)M_t + E_t \qquad (0 < \phi < 1),$$

where ϕ is the depreciation rate of CO2 (rate of removal from the atmosphere). Temperature H_t develops according to

$$H_{t+1} = \eta_0 + \eta_1 H_t + \eta_2 \log(M_{t+1}) \qquad (\eta_1 > 0,\ \eta_2 > 0).$$

In each period t, the fraction of GDP not spent on abatement or 'damage' is either consumed (C_t) or invested (I_t) along the budget constraint

$$(1 - \omega_t)d_t Y_t = C_t + I_t. \tag{3.1}$$

The temperature-impact function d_t depends only on temperature and satisfies $0 < d_t \leqslant \bar{d}_t$, where $\bar{d}_t$ represents the optimal temperature for the economy. Deviations from the optimal temperature cause damage. We specify d_t as

$$d_t = \frac{\bar{d}_t}{1 + \xi H_t^2} \qquad (\xi > 0).$$

For very high and very low temperatures d_t approaches zero. The optimal value of d_t occurs at $H_t = 0$ (the temperature in 1900, as in Nordhaus) when $d_t = \bar{d}_t$. Hence, 'net' output $d_t Y_t$ is a fraction, not of Y_t as in Nordhaus, but of $\bar{d}_t Y_t$, the output achievable under optimal climate conditions. A fraction ω_t

of $d_t Y_t$ is spent on abatement, and we specify the abatement cost fraction as

$$\omega_t = \psi_t \mu_t^\theta \qquad (\theta > 1).$$

If μ_t increases then so does ω_t, and a larger fraction of GDP will be spent on abatement. These equations capture the essence of the Nordhaus (2008) DICE model.

The model includes stock variables L_t, K_t, M_t, and H_t, fractions ω_t and μ_t, and scale variables A_t, d_t, σ_t, and ψ_t, all measured at the beginning of period t; and flow variables Y_t, C_t, I_t, and E_t, all measured in period t (not in year t). Notice that L_t is a stock, not a flow. As in Nordhaus (2008) one period is ten years. We choose the exogenous variables such that $L_t > 0$, $A_t > 0$, $\sigma_t > 0$, and $0 < \psi_t < 1$. The policy variables must satisfy

$$C_t \geqslant 0, \quad I_t \geqslant 0, \quad 0 \leqslant \mu_t \leqslant 1. \tag{3.2}$$

With these restrictions all variables will have the correct signs and all fractions will lie between zero and one.

3.2.2 Utility and welfare

Given a utility function U we define welfare in period t as

$$W_t = L_t U(C_t/L_t). \tag{3.3}$$

If the policy maker has an infinite horizon, then he/she will maximize total discounted welfare,

$$W = \sum_{t=0}^{\infty} \frac{W_t}{(1+\rho)^t} \qquad (0 < \rho < 1),$$

where ρ denotes the discount rate. Letting x denote per capita consumption, the utility function $U(x)$ is assumed to be defined and strictly concave for all $x > 0$. There are many such functions, but a popular choice is

$$U(x) = \frac{x^{1-\alpha} - 1}{1 - \alpha} \qquad (\alpha > 0), \tag{3.4}$$

where α denotes the elasticity of marginal utility of consumption. This is the so-called power function. If we define

$$\text{ARA}(x) = -\frac{U''(x)}{U'(x)}, \qquad \text{RRA}(x) = -\frac{xU''(x)}{U'(x)}, \tag{3.5}$$

then its coefficient of absolute risk aversion $\text{ARA}(x) = \alpha/x$ is decreasing and its coefficient of relative risk aversion $\text{RRA}(x) = \alpha$ is constant. The power function may be popular, but it does have drawbacks. In particular, we have $\text{RRA}(0) > 0$, which implies that the expected pricing kernel may not exist in the presence of heavy tails (Appendix 3.C). Later we shall therefore consider other utility functions as well.

The power function is bounded from below, but not from above when $0 < \alpha < 1$; and it is bounded from above, but not from below when $\alpha > 1$. When $\alpha = 1$ we have $U(x) = \log(x)$, which is unbounded from below and above. Many authors, including Nordhaus (2008), choose $\alpha = 2$. Also popular is $\alpha = 1$ (Kelly and Kolstad, 1999; Stern, 2007).

Our interest is in maximizing welfare W with respect to the policy bundles (C_t, I_t, μ_t) for $t = 0, 1, 2, \ldots$. In Table 3.1 we present the parameters and initial values used. These values are chosen such that our results closely resemble the results obtained by Nordhaus (2008), when applied within the same 60-period (600-year) DICE framework. We choose the exogenous variables L_t, A_t, σ_t, and ψ_t as in Nordhaus (2008), and we let $\bar{d}_t = 1$ and $\alpha = 2$.

Our GAMS code (http://center.uvt.nl/staff/magnus/catastrophe) then produces optimal values over sixty periods that are very close to the values obtained in Nordhaus, as shown in Table 3.2. Hence it appears that our simplified

Parameter	Value	Description
Endogenous stocks: initial levels		
K_0	137	Capital stock, begin of period 0
M_0	808.9	CO2 concentration, begin of period 0
H_0	0.731	Temperature, begin of period 0
Technology		
γ	0.30	Elasticity of capital in production function
δ	0.6513	Depreciation rate on capital, per decade
Pollution, damage, and abatement		
ϕ	0.0524	Depreciation rate on CO2 concentration, per decade
ξ	0.0028388	Quadratic term, temperature-impact function
θ	2.80	Exponent in abatement function
Temperature		
η_0	−5.9839	Constant term, temperature equation
η_1	0.7708	Previous period impact, temperature equation
η_2	0.9373	CO2 concentration impact, temperature equation
Discount rate		
ρ	0.1605	Welfare discount rate, per decade

Table 3.1: Parameter values for simplified DICE (SICE) model

	2005		2055		2105		2155	
	DICE	SICE	DICE	SICE	DICE	SICE	DICE	SICE
K	137	137	353	354	707	711	1317	1324
M	809	809	1048	988	1270	1233	1428	1430
H	0.7	0.7	1.8	1.5	2.7	2.4	3.3	3.2

Table 3.2: Comparison of stocks in Nordhaus (DICE) and our (SICE) models

version of the DICE model (hereafter, SICE = simplified DICE) works as the original version.

3.2.3 Uncertainty

So far we have ignored uncertainty. There is however much uncertainty in the economics of climate change (Manne and Richels, 1992; Nordhaus, 1994; Weitzman, 2009). There is model uncertainty, parameter uncertainty, and uncertainty about the possible reduction of parametric variability over time

(updating); see Kelly and Kolstad (1999) and Leach (2007). We model uncertainty through stochasticity. In the literature, stochasticity is typically introduced through the damage function (Roughgarden and Schneider, 1999; Mastrandrea and Schneider, 2004) or through a random shock in temperature (Kelly and Kolstad, 1999; Leach, 2007). We follow this literature by introducing stochasticity through the temperature-impact function d_t, more precisely through $\bar{d}_t$, the impact under optimal temperature. We are uncertain about the optimal temperature, because we are uncertain about the correctness of the functional form of d_t, about the values of the parameters, and about the underlying temperature equation. We capture these three sources of uncertainty by writing

$$\bar{d}_t = e^{-\tau^2/2}\, e^{\tau \epsilon_t},$$

where ϵ_t denotes a random error with mean zero and variance one. This implies that 'net GDP' is given by

$$d_t Y_t = \frac{e^{-\tau^2/2} Z_t}{1 + \xi H_t^2}, \qquad Z_t = A_t K_t^{\gamma} L_t^{1-\gamma} e^{\tau \epsilon_t}, \tag{3.6}$$

so that random noise enters the Cobb-Douglas production function in the usual 'linear' way when we write $\log(Z_t/L_t) = \log A_t + \gamma \log(K_t/L_t) + \tau \epsilon_t$.

If ϵ_t follows a normal distribution $\mathrm{N}(0,1)$, then the moments of $\bar{d}_t$ exist, and we have $\mathrm{E}(\bar{d}_t) = 1$ and $\mathrm{var}(\bar{d}_t) = e^{\tau^2} - 1$. Since the distribution of $\bar{d}_t$ is heavily skewed, its expectation is larger than its median, and hence more uncertainty (higher τ) implies more probability mass of $\bar{d}_t$ close to zero, and a higher probability of damage. If, however, we move only one step away from the normal distribution and assume that ϵ_t follows a Student distribution with *any* (finite) degrees of freedom, then the expectation is infinite (Geweke, 2001). (Heavy-tailed distributions (see Appendix 3.C for a formal definition) such as the Student distribution are natural in the context of extreme climate change.) This fact predicts that expected welfare may be very sensitive to distributional assumptions: random noise with finite moments (Student distribution) may turn into random variables without moments ($\bar{d}_t$, $d_t Y_t$).

3.2.4 Scrap values

If the policy maker has a T-period policy horizon, then we write welfare as

$$W = \sum_{t=0}^{T-1} \frac{L_t U(x_t)}{(1+\rho)^t} + \frac{1}{(1+\rho)^T} \sum_{t=0}^{\infty} \frac{L_{T+t} U(x_{T+t})}{(1+\rho)^t},$$

where $x_t = C_t/L_t$ denotes per capita consumption in period t. If $\{x_t^*\}$ denotes the optimal path for $\{x_t\}$, then we define the scrap value as

$$S_T = \sum_{t=0}^{\infty} \frac{L_{T+t} U(x_{T+t}^*)}{(1+\rho)^t}.$$

Maximizing W is then equivalent to maximizing

$$\sum_{t=0}^{T-1} \frac{L_t U(x_t)}{(1+\rho)^t} + \frac{S_T}{(1+\rho)^T}.$$

The scrap value S_T will depend on the state variables at time T, in particular K_T and M_T, and this functional relationship is the scrap value function: $S_T = S(K_t, M_T)$. If T is large we may ignore the scrap value S_T because of the large discount factor $(1+\rho)^T$. But if T is small, then we need to model S_T explicitly, thus emphasizing the fact that the policy maker has the double objective of maximizing discounted welfare over a finite number of periods T, while also leaving a reasonable economy for the next policy maker, based on the remaining capital stock and CO2 concentration.

The simplest approximation to S_T is the linear function

$$S_T = \nu_0 + \nu_1 K_T - \nu_2 M_T \qquad (\nu_1 > 0, \nu_2 > 0), \tag{3.7}$$

where ν_1 and ν_2 denote the scrap prices of capital and pollution at the beginning of period T. This scrap value function captures the idea that the next government will be happier if there is more capital and less pollution at the beginning of its policy period. But the linear scrap value function has

some problems. These are discussed in Chapter 5 where we also propose the non-linear scrap value function,

$$S_T = \nu_0 - \frac{\nu_1 K_0}{p}\left(\frac{K_T}{K_0}\right)^{-p} - \frac{\nu_2 M_0}{q}\left(\frac{M_T}{M_0}\right)^{q}, \tag{3.8}$$

where $\nu_1 > 0$, $\nu_2 > 0$, $p > 0$, and $q > 1$. This function is strictly concave, bounded from above, and approaching $-\infty$ when either $M_T \to \infty$ or $K_T \to 0$. In addition, it has the property that if we linearize $S(K_T, M_T)$ around (K_0, M_0) we find

$$S_T \approx \text{constant} + \nu_1 K_T - \nu_2 M_T,$$

so that ν_1 and ν_2 can be interpreted as scrap prices, just as in the linear case.

3.3 A two-period model with CRRA preferences

The simplest version of the model occurs when $T = 2$ in which case we have only two periods. We can write welfare in this case as

$$W = W(\mu_0, C_0, \mu_1, C_1, \epsilon_1) = W_0 + \frac{W_1}{1+\rho} + \frac{S_2}{(1+\rho)^2}.$$

The two-period model, which we will consider henceforth, captures the essence of our problem while remaining numerically tractable in the presence of uncertainty. In this section the utility function is given by $U(x) = 1 - 1/x$ (power utility with $\alpha = 2$), the random errors ϵ_t are generated by a normal $\mathrm{N}(0, 1)$ distribution, and the policy restrictions (3.2) are explicitly imposed, so that we maximize a restriction of expected welfare; see Appendix 3.A. Randomness results from d_1 only, because the temperature-impact d_0 at the beginning of period 0 is known to us (we set $\bar{d}_0 = 1$, equal to its expectation), and d_2 at the end of period 1 does not appear in the welfare function. Hence, the only source of randomness is caused by the error ϵ_1. The policy maker has to choose

the policy bundles (C_0, I_0, μ_0) at the beginning of period 0 and (C_1, I_1, μ_1) at the beginning of period 1 that will maximize expected welfare.

Parameter	Value	Description
Population		
L_0	6514	Population, begin of period 0
L_1	7130	Population, begin of period 1
Technology		
A_0	0.2722	Total factor productivity, begin of period 0
A_1	0.3000	Total factor productivity, begin of period 1
Pollution		
σ_0	0.1342	CO2 emissions-to-output ratio, period 0
σ_1	0.1253	CO2 emissions-to-output ratio, period 1
Abatement		
ψ_0	0.0561	Coefficient in abatement function, period 0
ψ_1	0.0511	Coefficient in abatement function, period 1

Table 3.3: Exogenous variables in the two-period model

We need values for the exogenous variables L_t, A_t, σ_t, and ψ_t. These are given in Table 3.3. Since a linear scrap value function is not realistic, because the combination of a half-bounded utility function and an unbounded scrap value function is theoretically not possible, we consider the non-linear scrap value function proposed in (3.8) with

$$\nu_1 = 286.15, \quad \nu_2 = 3.60, \quad p = 0.20, \quad q = 2.0.$$

Finally, we need sensible values for the uncertainty parameter τ. The stochasticity, as given in (3.6), captures uncertainty about GDP that is due to uncertainty about climate change. Historical variation in GDP may therefore serve as an initial upper bound proxy for τ. Barro (2009) calibrates the standard deviation of log GDP to a value of 0.02 on an annual basis. Over a 10-year horizon this would correspond to about 0.06, under normality. Barro, however, only considers rich (OECD) countries, which means that for our purposes this value needs to be scaled up.

In Figure 3.2 we plot the density of $\bar{d}_1$ for three values of τ: 0.1, 0.3, and 0.7, both when ϵ_1 follows a $\mathrm{N}(0,1)$ distribution (solid line) and when $\epsilon_1 = \sqrt{4/5}\,t$,

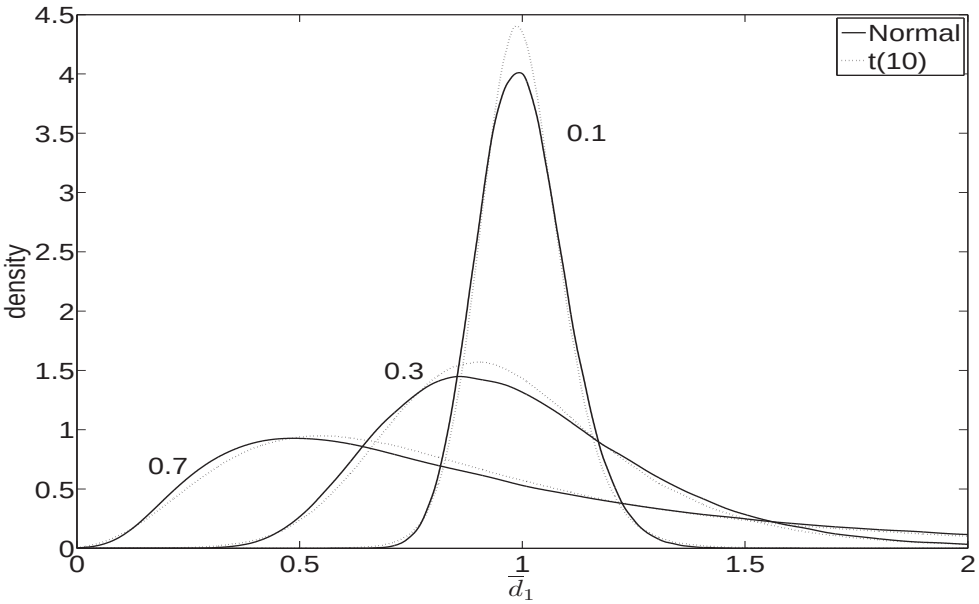


Figure 3.2: Density of $\bar{d}_1$ for $\tau = 0.1$, 0.3, and 0.7

where t follows a Student distribution with 10 degrees of freedom. Notice that $\mathrm{E}(\epsilon_1) = 0$ and $\mathrm{var}(\epsilon_1) = 1$ in both cases. When $\tau = 0.1$, we see that almost 100% of the distribution of $\bar{d}_1$ lies in the interval $(0.5, 2.0)$, both for the $\mathrm{N}(0, 1)$ distribution and for the $t(10)$ distribution. When $\tau = 0.3$, 97.8% (97.2% for the Student distribution) lies in the interval $(0.5, 2.0)$; and, when $\tau = 0.7$, only 64.9% (67.2% for the Student distribution) lies in this interval. We conclude that $\tau = 0.7$ may serve as a credible upper bound for the uncertainty range, and hence we report our results for $\tau = 0.0$, 0.3, and 0.7.

Realizing that at the beginning of period 1 the temperature-impact d_1 is observed based on the realization of ϵ_1, the policy maker will maximize expected welfare in three steps as follows. First, he/she maximizes welfare $W = W(\mu_0, C_0, \mu_1, C_1, \epsilon_1)$ with respect to (μ_1, C_1) conditional on (μ_0, C_0, ϵ_1) and under the restriction (3.2). This gives (μ_1^*, C_1^*) and concentrated welfare

$$W^*(\mu_0, C_0, \epsilon_1) = W(\mu_0, C_0, \mu_1^*, C_1^*, \epsilon_1).$$

Then the expectation $\overline{W}(\mu_0, C_0) = \mathrm{E}\left(W^*(\mu_0, C_0, \epsilon_1)\right)$ is computed, if it exists. Finally, $\overline{W}$ is maximized with respect to (μ_0, C_0). With the parameter values

τ	0.0	0.3	0.7		0.0	0.3	0.7
Policy instruments				*Stocks*			
μ_0	0.0933	0.0920	0.0874	K_1	172.49	178.66	204.33
μ_1	0.1137	0.1141	0.1142	K_2	211.66	219.32	255.70
C_0	431.07	424.90	399.24	M_1	834.25	834.34	834.69
C_1	547.20	549.08	550.63	M_2	868.31	869.18	872.74
I_0	124.72	130.89	156.56	H_1	0.8843	0.8844	0.8848
I_1	151.51	157.03	184.45	H_2	1.0400	1.0411	1.0452

Table 3.4: Power utility under normality

and exogenous variables given in Tables 3.1 and 3.3, and the four parameter values in the non-linear scrap value function given above, we obtain the results presented in Table 3.4. We note here and in subsequent tables that $Y_0 = 556.67$ and $d_0 = 0.9985$ are constant over different scenarios and functions, and that the values of μ_0, C_0, I_0, E_0, ω_0, K_1, M_1, and H_1 are optimal values. In contrast, μ_1, C_1, I_1, Y_1, E_1, ω_1, d_1, K_2, M_2, and H_2 are optimal *functions* of ϵ_1. What we present in the tables are their expectations.

For $\tau = 0$ there is no uncertainty. For $\tau > 0$ there is uncertainty, and all policy variables are affected when τ increases. More uncertainty results in less abatement, less consumption, and more investment in period 0, and to more abatement, consumption, and investment in period 1. In period 1, the changes in abatement and consumption are negligible. The increase in I_0 with τ can be explained by precautionary savings. The restriction on I_1 can be viewed as a penalty for negative investment. To avoid this penalty, the policy maker can increase the budget in period 1 by investing more in period 0. As the amount of uncertainty increases, the probability of negative investment increases, *ceteris paribus*. In response, the policy maker increases investment at the expense of abatement and consumption in period 0. The increase in I_0 leads to higher output in period 1, which explains the increases in I_1, K_2, M_2, and H_2. The decrease in μ_0 leads to higher emissions in period 0, and increases carbon concentration and temperature in period 1. An additional reason why investment in period 1 increases with uncertainty is that positive shocks translate into possibly unlimited upward shocks in I_1, but negative shocks will never cause I_1 to drop below zero.

We need to show that the expectation of welfare exists for power utility. The following proposition states not only this but also that, if we move one step away from normality and assume a Student distribution with any finite degrees of freedom, then the expectation does not exist.

Proposition 3.1. With power utility, expected welfare exists under normality but not under a Student distribution.

The proof of Proposition 3.1 is in Appendix 3.B. It follows that the much-used power utility function is inconsistent with expected utility theory with heavy tails, not because utility theory itself is at fault but because power utility is inappropriate when tails are heavy.

3.4 Catastrophic risk and compatibility

3.4.1 Expected utility and catastrophic risk

Since the axiomatization of expected utility (EU) by Von Neumann and Morgenstern (1944) and Savage (1954) numerous objections have been raised against it. Most of these relate to empirical evidence that the behavior of agents under risk and uncertainty does not agree with EU. Indeed there is much evidence that for descriptive applications the Von Neumann and Morgenstern axioms are violated systematically. Motivated by such empirical evidence, various alternative theories have emerged, usually coined 'non-expected utility' theories; see Sugden (1997) for a review. Starting with the Allais paradox in the 1950s, problems involving low-probability high-impact outcomes have played a central role in non-expected utility. This is particularly, but not exclusively, evident in the rank-dependent class of models such as Kahneman and Tversky's cumulative prospect theory.

One option in our context would be to dismiss EU and replace it by a non-expected utility theory. While this could conceivably, at least partially, solve the problem of maximizing EU for catastrophic risks (in particular, the existence of moments), it might also aggravate the problems. Non-expected

utility theories, most notably prospect theory, account for the fact that people are limited in their ability to comprehend and evaluate extreme probabilities, so that highly unlikely events are either ignored or overweighted. But, for normative purposes, it is dangerous to ignore or overweight highly unlikely events, and policy makers should choose a framework where this is avoided.

Despite important developments in non-expected utility theory, EU remains the dominant *normative* decision theory (Broome, 1991; Sims, 2001; Dhami and Al-Nowaihi, 2010), and the current paper stays within the framework of EU. Our results presented below corroborate the fact that expected utility theory may reliably provide normatively appealing results, also in the presence of catastrophic risks. Nevertheless, one may legitimately question whether EU is the appropriate normative theory for decision making under catastrophic risks and continue a search for better theories; see also Chichilnisky (2000).

In Appendix 3.C we derive necessary and sufficient conditions on the utility function to ensure that expected utility and expected marginal utility (hence also the expected pricing kernel) are finite, also in the presence of heavy tails. These results are generally applicable to standard multi-period welfare maximization problems. This is important, because if the expected pricing kernel is infinite, then the amount of consumption in period 0 which the representative agent is willing to give up in order to obtain one additional certain unit of consumption in period 1 is infinite. This is not credible and, following our discussion in the Introduction, we argue that such a view of life is unreasonable and has extreme irrational implications in our setting. The price we are willing to pay to avoid a global economy-climate catastrophe is finite.

Propositions 3.3 and 3.4 underline the importance of a compatible specification of the concavity index, especially in the presence of catastrophic risk. We note in this context that the concavity index is to be specified as input, and need not be estimated, though it may be learned, that is, implicitly elicited (see Section 3.5 below). In what follows, we will provide translations of the concavity index parameters into quantities that allow a more straightforward interpretation. These translations can then serve as handles to set the input

parameters. Quantities with a relatively simple meaning will be the key to the process of learning about the concavity index parameters.

3.4.2 Compatibility: Non-normality and Burr utility

Motivated by the conditions derived in Appendix 3.C and by the fundamental insight that the economic model and the statistical model must be compatible, and because we wish to leave distributional assumptions unrestricted at this stage, we consider two bounded utility functions: the exponential function and the 'Burr' function. (Other choices are permitted but may require restrictions on distributional assumptions.) The exponential utility function is given by

$$U(x) = 1 - e^{-\beta x} \qquad (\beta > 0) \tag{3.9}$$

with $\text{ARA}(x) = \beta$ and $\text{RRA}(x) = \beta x$, and the Burr utility function by

$$U(x) = 1 - \left(\frac{\lambda}{x+\lambda}\right)^{k} \qquad (k > 0,\, \lambda > 0) \tag{3.10}$$

with $\text{ARA}(x) = (k+1)/(x+\lambda)$ and $\text{RRA}(x) = (k+1)x/(x+\lambda)$. Both functions are members of the HARA class of utility functions. The Burr function, based on Burr (1942) and Burr and Cislak (1968), was proposed in Chapter 4, where it is also shown that this function is particularly appropriate as an approximation to a bounded utility function and enjoys a combination of appealing properties especially relevant in heavy-tailed risk analysis. This is exemplified in Figure 3.3, where we plot RRA and ARA for the power function ($\alpha = 2$), the exponential function ($\beta = 25$), and the Burr function ($k = 1.5$, $\lambda = 0.02$). The parameter choice is determined by the point x^*, where we want the three functions to be close. Suppose we want the functions to be close at $x^* = 0.08$, which is approximately the value of C_0/L_0 and C_1/L_1. Then, given that $\alpha = 2$, we find $\beta = 2/x^* = 25$, and, for any $k > 1$, $\lambda = (k-1)x^*/2$. Intertemporal preferences are jointly determined by the RRA parameters (α, β, k, λ) and the discount rate ρ. In our case we keep

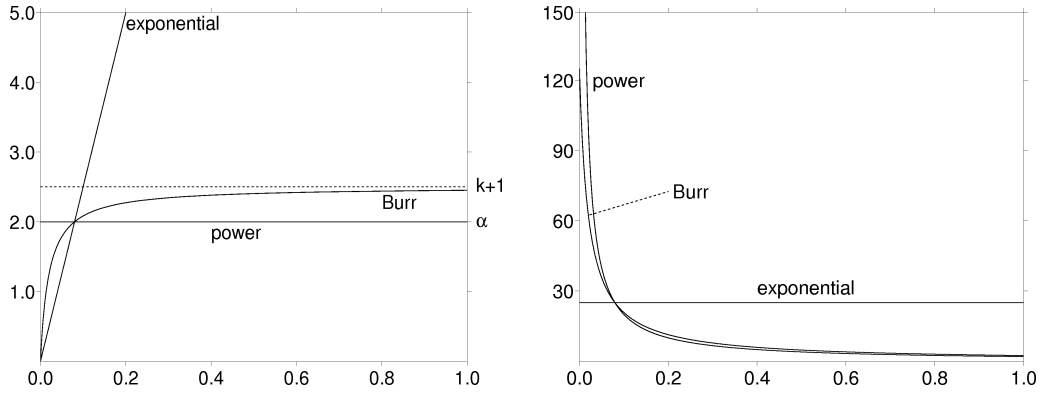


Figure 3.3: RRA (left) and ARA (right) for three utility functions

ρ constant and choose the RRA-ARA parameters appropriately according to the above closeness criterium.

The power function has RRA(0) > 0 and ARA(0) $= \infty$, while the RRA in the exponential function is unbounded for large x. In contrast, the RRA in the Burr function is bounded between 0 and $k+1$, it satisfies RRA(0) $= 0$, and ARA(0) is finite (125 in the figure). Notice that the fact that RRA(0) $= 0$ (as is the case for the exponential and the Burr utility functions) does *not* imply that the representative agent is risk-neutral at $x = 0$. For example, we have ARA(0) $= \beta$ for the exponential function and ARA(0) $= (k+1)/\lambda$ for the Burr function. Also, the derivative RRA$'$ is zero for the power function, constant for the exponential function, and monotonically decreasing from $(k+1)/\lambda$ to zero for the Burr function. Hence the slope of the Burr function at $x = 0$ is finite.

The Burr function is attractive because it lies in-between the power and exponential functions. It exhibits exponential-like features when x is close to zero (with RRA(0) $= 0$, ARA(0) $< \infty$, and $U(0) > -\infty$), hence satisfies the compatibility conditions derived in Appendix 3.C; and power-like features in the middle and on the right-side of the distribution. Relative concepts are useful away from zero but not close to zero, and this is why power utility does not work well in near-catastrophe scenarios. Indeed, in the insurance literature exponential utility is typically used (Gerber, 1979, Chapter 5).

Using GAMS and without uncertainty, we maximize welfare over sixty periods (600 years) for both the exponential and Burr utility functions, and

	2005		2055		2105		2155	
	Expo	Burr	Expo	Burr	Expo	Burr	Expo	Burr
K	137	137	286	343	388	666	456	1220
M	809	809	1012	993	1328	1258	1727	1512
H	0.7	0.7	1.6	1.5	2.6	2.5	3.7	3.3

Table 3.5: Comparison of stocks in Exponential and Burr models

a selection of the resulting optimal values is shown in Table 3.5. When we compare the results with those in Table 3.2, we see that the optimal stock values from the Burr function closely resemble the optimal stock values from the power function, but not those from the exponential function. In contrast to power and Burr, where RRA flattens out, the RRA for the exponential distribution continues to increase (Figure 3.3), and hence the growth rate of marginal utility continues to increase as well. As x increases, consumption will therefore increase, and investment and abatement will decrease. As a result, C/Y is relatively large for exponential utility. The low growth rate of capital (for exponential utility) leads to a low growth rate of output. However, since more consumption leads to less abatement, the growth rate of CO2 concentration is high even when the amount of production is low. Consequently, M and H are high compared to power and Burr. When $x < x^*$, RRA (Burr) is close to RRA (exponential), so that more is consumed and less invested when the Burr function is used instead of the power function. But when $x > x^*$, RRA (Burr) is close to RRA (power). The optimal path of K is slightly lower and the optimal paths of M and H are slightly higher for Burr than for power utility.

The scrap value function for both utility functions, developed in Chapter 5, is defined as

$$S_T = \nu_0 - \nu_1\zeta_1\left(1 + K_T/\lambda_1\right)^{-p} + \nu_2\zeta_2\left(1 + (M_T/\lambda_2)^c\right)^{-q} \tag{3.11}$$

with

$$\zeta_1 = \frac{\lambda_1}{p}\left(1 + K_0/\lambda_1\right)^{p+1}, \qquad \zeta_2 = \frac{\lambda_2}{cq} \cdot \frac{\left(1 + (M_0/\lambda_2)^c\right)^{q+1}}{(M_0/\lambda_2)^{c-1}},$$

where $p > 0$, $q > 0$, $c > 1$, $\lambda_1 > 0$, and $\lambda_2 > 0$, and we have again normalized $\nu_1 > 0$ and $\nu_2 > 0$ such that $\partial S_T/\partial K = \nu_1$ at $K = K_0$ and $\partial S_T/\partial M = -\nu_2$ at $M = M_0$.

		Normal		Student(10)	
τ	0.0	0.3	0.7	0.3	0.7
Policy instruments					
μ_0	0.1175	0.1166	0.1135	0.1166	0.1135
μ_1	0.1473	0.1515	0.1697	0.1516	0.1700
C_0	428.74	425.43	413.86	425.41	413.52
C_1	551.45	584.64	527.76	∞	∞
I_0	127.00	130.32	141.90	130.35	142.24
I_1	149.95	156.68	190.03	∞	∞
Stocks					
K_1	174.78	178.10	189.67	178.12	190.01
K_2	210.90	218.78	256.16	∞	∞
M_1	832.44	832.51	832.74	832.51	832.74
M_2	863.94	864.06	864.07	864.06	864.09
H_1	0.8823	0.8824	0.8826	0.8824	0.8826
H_2	1.0337	1.0339	1.0341	1.0339	1.0341

Table 3.6: Exponential utility: Normal versus Student(10)

The values of the calibrated parameters for the scrap value functions in the case of exponential utility are $\nu_1 = 8.0282$, $\nu_2 = 0.1487$, and

$$p = 0.85, \quad q = 0.55, \quad c = 1.65, \quad \lambda_1 = 0.0936, \quad \lambda_2 = 0.0415.$$

The optimal values of the policy and other variables obtained from maximizing expected welfare are presented in Table 3.6. In contrast to Table 3.5, the results in Table 3.6 (and Table 3.7) allow for uncertainty, consider the short run (two periods) rather than the long run (sixty periods), and also take scrap values into account. Since exponential utility is calibrated to be close to power utility at $x = x^*$, the results for the two utility functions do not differ greatly. This is especially true for $\tau = 0$, where only the abatement fraction μ is higher for exponential utility, and therefore temperature H is lower.

When τ increases, I_0 increases less and I_1 increases more for exponential than for power. Moreover, as the uncertainty parameter τ increases, M_2 does not change much in the exponential case, while it increases in the power case. The effect of uncertainty on the marginal scrap values is therefore larger in the exponential case than in the power case. As in Table 3.4, more uncertainty results in less abatement, less consumption, and more investment in period 0, and to more abatement, and investment in period 1.

Suppose now that the underlying distribution has heavier tails: Student instead of normal. We have normalized the variance of the Student distribution to be one, and therefore the first three moments of ϵ_1 are the same as under normality. The kurtosis, however, is now slightly higher: 4 instead of 3 (assuming 10 degrees of freedom). Under power utility, expected welfare does not exist any more. But under bounded utility, expected welfare always exists. The effect of the excess kurtosis on the optimal values is relatively small. It is important to realize that, while the Student distribution features excess kurtosis, it remains quite close to the normal distribution (see also Figure 3.2). Hence it would be unreasonable if a 'small' change in distributional assumptions would lead to a large possibly 'discontinuous' change in optimal policies.

All variables move in the same direction as before when τ increases. Notice that some variables (C_1, I_1, and K_2) have infinite expectations even though expected welfare is finite. This is no surprise because these variables are unbounded and depend on $\bar{d}_1 = e^{-\tau^2/2}e^{\tau\epsilon_1}$. When ϵ_1 follows a Student distribution, $\mathrm{E}(\bar{d}_1) = \infty$ and this property carries over to the other three variables.

Let us now consider a second bounded utility function, the appealing Burr function. For Burr utility the following parameter values were calibrated for the scrap value functions: $\nu_1 = 2.8884$, $\nu_2 = 0.0366$, and

$$p = 0.6, \quad q = 0.5, \quad c = 1.5, \quad \lambda_1 = 0.1631, \quad \lambda_2 = 0.0502.$$

The optimal values are presented in Table 3.7. In view of Figure 3.3 we would expect that Burr and power are relatively close in the observed data range.

		Normal		Student(10)	
τ	0.0	0.3	0.7	0.3	0.7
Policy instruments					
μ_0	0.0924	0.0910	0.0859	0.0910	0.0861
μ_1	0.1124	0.1135	0.1175	0.1135	0.1175
C_0	430.76	424.31	399.97	424.33	400.67
C_1	548.53	552.59	563.14	∞	∞
I_0	125.03	131.48	155.83	131.46	155.12
I_1	150.56	154.23	171.09	∞	∞
Stocks					
K_1	172.80	179.25	203.60	179.23	202.89
K_2	210.81	216.73	242.08	∞	∞
M_1	834.32	834.42	834.80	834.42	834.79
M_2	868.53	869.39	872.45	869.38	872.36
H_1	0.8844	0.8845	0.8850	0.8845	0.8849
H_2	1.0403	1.0413	1.0450	1.0413	1.0449

Table 3.7: Burr utility: Normal versus Student(10)

This is indeed the case as a comparison of Tables 3.7 and 3.4 reveals. There is little difference between the two tables in the case of no uncertainty, and also when τ increases. The effect of excess kurtosis is again small, as it should be.

The important difference between power and Burr utility is not revealed in our typically observed data. It is only revealed when low levels of per capita consumption become relevant, that is, in near-catastrophe cases. This is clarified in Figure 3.4, where we present μ_1 as a function of ϵ_1 for $\tau = 0.3$. The expected value of μ_1 is 0.1141 for power utility under normality (Table 3.4), and 0.1135 for Burr utility under either normality or Student(10) (Table 3.7). This is not very different. But for values of ϵ_1 further away from 0 the difference is large.

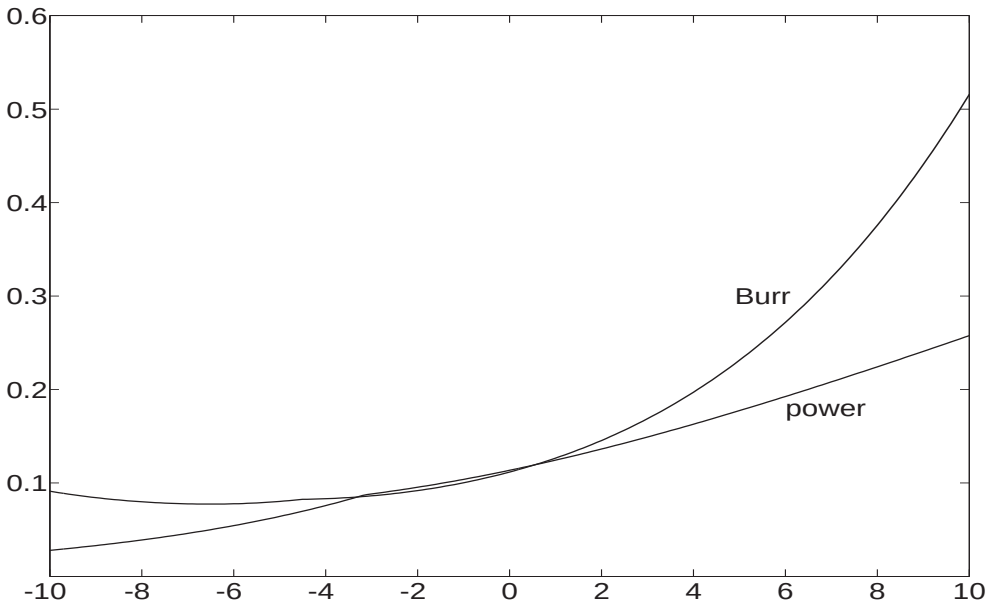


Figure 3.4: μ_1 as a function of ϵ_1: Burr versus power utility

3.5 Agreement and robustness

3.5.1 Learning and agreement

To complete the model we need to specify our input parameters. We show how this can be achieved in a process towards agreement, using experts' priors. The key to this learning and agreement process is the translation of model parameters that are relatively difficult to interpret into quantities that allow a more straightforward interpretation. The process is applicable not only in the current context of extreme climate change, but also in many other policy making settings involving catastrophic risks.

Our parameters cannot be estimated using conventional methods and historical data, but experts will have prior ideas about these parameters. Different experts will have different priors. Model output can be generated on the basis of various priors. Then, in an iterative procedure, one learns about the parameter values from experts' opinions and model output, and an agreeable intersection of model parameters may be reached.

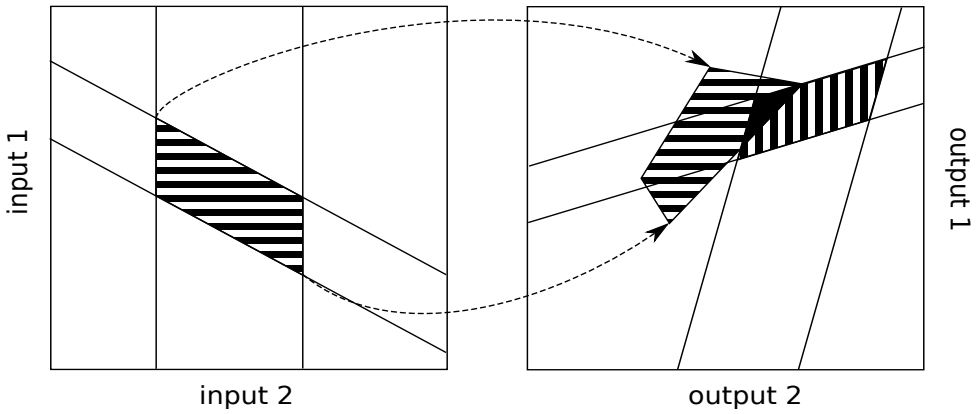


Figure 3.5: The decision making process

This process is illustrated in Figure 3.5. In the left panel, we visualize the contributions of two experts. One expert states that the value of input 2 should be bounded as indicated by the two vertical lines. The other expert provides a lower and upper bound for the value of input 1, depending on the value of input 2. The horizontally-shaded area gives the combinations of inputs that are acceptable to both experts. The right panel is more complicated. We first visualize the contributions of two policy makers regarding two output variables. This is the vertically-shaded area, giving the combinations of outputs that are acceptable to both policy makers. Next we map the left panel onto the right panel. For every acceptable combination of inputs the model provides one combination of outputs, that is, one point in the right panel. The horizontally-shaded area in the right panel is the image of the horizontally-shaded area in the left panel. We now have two areas in the right panel: the vertically-shaded area and the horizontally-shaded area. If the two areas do not intersect, then the experts and policy makers must adjust their priors in an iterative process of learning. Once the areas do intersect, agreement is possible. The black triangle then contains all points for which both inputs and outputs are acceptable. Agreement must be reached on the three policy variables (μ_0, C_0, I_0), and we recall that expected welfare is maximized in three steps as described in Section 3.3, yielding the optimal policy (μ_0^*, C_0^*, I_0^*).

Our analysis requires prior beliefs about various inputs, in particular: form of the utility function (Burr or otherwise), degree of risk-aversion (k, λ), discount rate (ρ), form of the distribution (Student or otherwise), and volatility (τ). If agreement is to be reached, then the policy makers must be willing to adjust their individual priors on each of these inputs, based on the experts' opinions and the generated output.

Since extreme outcomes matter, the normal distribution is not appropriate. We want a distribution which allows heavier tails, such as the Student distribution. Given our treatment of stochasticity, power utility is not compatible with the Student distribution, because the required expectations don't exist. Also, exponential utility has the disadvantage that RRA increases without bound. Burr utility provides a useful compromise: it exhibits exponential-like features when per capita consumption is small, and power-like features otherwise. Let us then confine ourselves to Burr utility, assume that ϵ_1 follows a Student distribution, and take the following parameter values as our benchmark:

$$k = 1.5, \quad \lambda = 0.02 \quad \tau = 0.3, \quad \text{df} = 10, \quad \rho = 0.1605.$$

Note that the value of λ is linked to k through $\lambda = 0.04(k-1)$, as explained in Section 3.4. The symbol df denotes the degrees of freedom in the Student distribution, and the discount rate of 0.1605 per decade corresponds to an annual discount rate of 0.015.

Our benchmark is column a in Table 3.8. The model outputs are within credible bounds: policy variables at the beginning of period 0 (μ_0, C_0, I_0); stock variables at the beginning of period 1 (K_1 and also $M_1 = 834.42$ and $H_1 = 0.8845$); and expectations (E(μ_1), E(H_2), and also E(M_2) = 869.38). If we consider temperature H_2 as a function of ϵ_1 we find relatively low volatility in comparison to the confidence intervals proposed by the IPCC (2007, Chapter 10). The reason for this is twofold. First, the IPCC determines confidence intervals by considering multiple deterministic climate models, not a single stochastic one as we do. Second and more importantly, the IPCC confidence intervals are based on non-mitigation scenarios, while our model takes policy

	Agreement				*Robustness*		
	a	*b*	*c*	*d*	*e*	*f*	*g*
Parameter values							
τ	0.3	0.3	0.5	0.3	0.5	0.7	0.5
df	10	25	10	10	25	10	10
k	1.5	1.5	1.5	2.0	1.5	1.5	2.0
Policy instruments, beginning of period 0							
μ_0	0.0910	0.0910	0.0888	0.1192	0.0887	0.0861	0.1163
C_0	424.33	424.31	413.71	438.01	413.50	400.67	427.56
I_0	131.46	131.47	142.08	117.73	142.29	155.12	128.19
Capital stock and expectations							
K_1	179.23	179.25	189.86	165.50	190.06	202.89	175.96
μ_1	0.1135	0.1135	0.1154	0.1604	0.1154	0.1175	0.1655
H_2	1.0413	1.0413	1.0429	1.0309	1.0430	1.0449	1.0323
Probabilities of catastrophe π_ℓ							
π_a	5.0E−03	3.3E−03	5.2E−02	5.1E−03	5.3E−02	1.4E−01	5.2E−02
π_b	2.3E−05	5.9E−07	1.4E−03	2.6E−05	5.5E−04	1.2E−02	1.5E−03
π_c	2.5E−07	5.0E−11	2.8E−05	2.6E−07	8.2E−07	4.9E−04	3.0E−05
Values of statistical subsistence $V_\ell = VSS_\ell/C_0$							
V_a	2.8E+01	2.9E+01	4.8E+00	2.4E+01	4.1E+00	2.9E+00	4.1E+00
V_b	1.3E+04	2.8E+05	3.1E+02	1.1E+04	5.3E+02	5.2E+01	2.6E+02
V_c	1.9E+06	4.2E+09	2.0E+04	1.6E+06	3.6E+05	1.4E+03	1.7E+04

Table 3.8: Parameter calibration based on Burr utility and Student distribution

effects into account. For both reasons, the volatility in temperature found by the IPCC is higher than what we find.

3.5.2 Probability of catastrophe and value of statistical subsistence

In addition to the 'direct' outputs of our model we also have 'derived' outputs, in particular the probability of catastrophe. These derived outputs are functions of the direct outputs and they represent important policy variables on which prior information is available. Hence, they also require agreement.

We propose to define catastrophe as the event $C_1^* \leqslant \underline{C}$ for some given value $\underline{C} > 0$. The probability of catastrophe is then given by $\pi = \Pr(C_1^* \leqslant \underline{C})$. We

shall consider three different values of $\underline{C}$: $\underline{C}_a$, $\underline{C}_b$, and $\underline{C}_c$, corresponding to three levels of catastrophe, labeled A, B, and C. Catastrophe A occurs when 20% of the world population live in extreme poverty, and catastrophes B and C occur when 50% and 80% of the world population live in extreme poverty, respectively. (The definitions and priors proposed in this subsection are based on background material provided in Ikefuji *et al.* (2010) and available at the project's website `http://center.uvt.nl/staff/magnus/catastrophe`.)

We must agree on acceptable values for the probability π of catastrophe. We have studied acceptable risks in various situations, and we conclude that an acceptable probability for an economy-climate catastrophe in the next 10-year period is in the range 10^{-5}–10^{-6}. Given the definition of catastrophe we propose: $\pi_a = 0.1$, $\pi_b = 0.001$, and $\pi_c = 0.00001$ as reasonable values. There are of course other definitions of catastrophe. Barro and Ursúa (2008) define catastrophe as a peak-to-trough fall in per capita GDP of at least 15%, and find that the probability of this happening is approximately $\pi = 0.017$ per year. This does not relate directly to our π values, because GDP is not the same as consumption, Barro and Ursúa consider one year while we work with 10-year intervals, and, most importantly, because they only consider 21 'rich' countries. So, the numbers are difficult to compare. A situation where 20% of the world live in extreme poverty has in fact occurred before. The percentage of the world's population living in extreme poverty has halved since 1981. So a probability of $\pi_a = 0.1$ seems reasonable.

In the benchmark model we find $\pi_a = 0.005$, $\pi_b = 0.00002$, and $\pi_c = 0.0000003$, which is much lower than the acceptable values. Given the associated costs, it seems unnatural that policies would be chosen that mitigate the probability of a global economy-climate catastrophe far beyond acceptable levels. What can we do about this? One possibility is to make the tails heavier or lighter, that is, to adjust the degrees of freedom. If we set df = 25 then π becomes even smaller. In general, π becomes smaller as the tails become lighter (df increases), as one would expect. For df = ∞ (the normal distribution) we find $\pi_a = 2.3\text{E}{-}03$, $\pi_b = 5.3\text{E}{-}10$, and $\pi_c = 1.5\text{E}{-}24$. Interestingly, the policy variables are hardly affected (column b), not even when df = 200

or df $= \infty$. If we set df $= 3$, which is the minimum value where var(ϵ_1) exists, then $\pi_a = 0.008$, a little higher than for df $= 10$, but not enough. So, adjusting the degrees of freedom hardly changes the results.

Perhaps the fact that the heaviness of the tail (degrees of freedom) has little effect on the optimal policy is caused by the Burr utility function. Maybe this function does not distinguish well between light and heavy tails? In fact, this is not so. It follows from Figure 3.2 (and Section 3.4.2) that τ has much more impact than df. Hence the Burr function does distinguish between light and heavy tails.

Perhaps we should then adjust the value of τ. In our benchmark we set $\tau = 0.3$ as a reasonable starting point. We could revise τ upwards. We argued in Section 3.3 and Figure 3.2 that $\tau = 0.7$ is an upper bound to the volatility. Let us therefore consider the case $\tau = 0.5$. A larger value of τ means more volatility and hence one would expect less consumption and more investment. This is indeed what happens (column *c*). Also, the probabilities are affected and are now much closer to our prior ideas.

We can also adjust the curvature k (and λ). If k increases, then agents become less risk-averse and, as expected, there is more consumption and less investment (column *d*). The probabilities are not much different from our benchmark in *a*, but the values of μ_0 and μ_1 are very high and the capital stock accumulation rate is only 1.9% per year, which is too low.

Finally, we could adjust the discount rate ρ. This is an important issue (see, for example, Gollier, 2002, 2008, and the references therein), with possibly significant (yet not 'discontinuous') impact on the optimal policies. It is, however, beyond the scope of this study.

Based on these comparisons it seems that policy *c* should be recommended. There is, however, one other derived output which is often discussed, namely the value of statistical life. If we agree on the definition of catastrophe, then we can also define the 'value of a statistical subsistence' (VSS) as the amount of consumption in period 0 that the government is willing to trade off in order to change the probability of catastrophe; see Ikefuji *et al.* (2010). The VSS is similar to the value of statistical life (VSL), except that it refers to the condition of just having enough food to stay alive (more than \$1/day) rather

than to life. We define (Ikefuji *et al.*, 2010)

$$\text{VSS} = \frac{1}{\partial\pi(C_0)/\partial C_0}$$

evaluated at $C_0 = C_0^*$. We propose $\text{VSS}_a = C_0$, $\text{VSS}_b = 10C_0$, and $\text{VSS}_c = 100C_0$ as reasonable orders of magnitude. We first need to establish that the VSS exists. A little algebra shows that the VSS can be expressed as $\text{VSS} = \Gamma_1/\Gamma_2$ with

$$\Gamma_1 = \text{E}\left(W_1(C_0^*, C_1^*) \,|\, C_1^* > \underline{C}\right) - \text{E}\left(W_1(C_0^*, C_1^*) \,|\, C_1^* \leqslant \underline{C}\right)$$

and

$$\begin{aligned}\Gamma_2 = {} & (1+\rho)\partial W_0(C_0)/\partial C_0 + (1-\pi)\partial\,\text{E}\left(W_1(C_0, C_1^*) \,|\, C_1^* > \underline{C}\right)/\partial C_0 \\ & + \pi\partial\,\text{E}\left(W_1(C_0, C_1^*) \,|\, C_1^* \leqslant \underline{C}\right)/\partial C_0 + \frac{1}{1+\rho}\partial\,\text{E}\left(S_2(C_0, C_1^*)\right)/\partial C_0,\end{aligned}$$

and where π and all partial derivatives are evaluated at $C_0 = C_0^*$. Under Burr utility all expectations exist.

The VSS (and the VSL) is a difficult concept to measure, and the VSS priors may be unreliable. As such it should not carry too much weight as a derived output. Still we notice that the VSSs of our preferred policy *c* are much closer to our reasonable values than the VSSs in columns *a*, *b*, and *d*. The VSSs in column *c* are quite high though. Apparently society is willing to sacrifice $5C_0$ to avoid catastrophe *A* and even $20{,}000C_0$ to avoid catastrophe *C*. Perhaps our 'reasonable' values are too small. In fact, this is a well-known problem. Weitzman (2009) discusses it and he also mentions large values for the VSL without it being clear what the consequences are.

3.5.3 Robustness

If we believe that column *c* is the best, then we should do some further robustness checks, starting from column *c* rather than column *a*. We have done

extensive robustness checks and some representative results of this analysis is reported in columns e–g of Table 3.8. If we adjust the degrees of freedom (column e), then not much happens. There is little to choose between columns c and e. The optimal policy (μ_0^*, C_0^*, I_0^*) is hardly affected, which is a good thing, because it means that our policy is not too sensitive to changes in the heaviness of the tail (degrees of freedom). In column f we consider $\tau = 0.7$. Here the probabilities of catastrophe seem to be too large. For example, we have $\pi_c = 0.0005$ and it is doubtful if the government would find this acceptable. The choice of volatility τ does, however, affect the policy, and hence is important. We see that economics and statistics are bound together and difficult to separate. In column g we adjust the curvature of the Burr utility function. The probabilities are hardly affected but there will be more consumption, less investment, and in particular more (perhaps too much) abatement. On the basis of these and other robustness checks we conclude that policy c is robust against small changes in the underlying assumptions and parameter values.

3.5.4 Weitzman's dismal theorem

The previous discussion is closely related to an important debate initiated by Weitzman (2009). In a highly stylized setting, Weitzman notices that heavy-tailed uncertainty and power utility are incompatible, as this combination of uncertainty and preferences implies an infinite pricing kernel. In order to avoid this, Weitzman introduces a lower bound on consumption. He then shows that this lower bound is related to a parameter that resembles the value of a statistical life, and proves that the pricing kernel approaches infinity as the value of this parameter approaches infinity (the 'dismal theorem'). Weitzman further argues that this 'VSL-like' parameter is hard to know, and interprets this result as follows:

> "... reasonable attempts to constrict the length or the heaviness of the 'bad' tail (or to modify the utility function) still can leave us with uncomfortably big numbers whose exact value depends non-robustly upon artificial constraints or parameters that we really do not understand." (Weitzman, 2009, p. 11)

We agree with Weitzman that incompatible combinations of utility functions and distribution functions exist, in the sense that the pricing kernel or other important policy variables become infinite. In fact we derive necessary and sufficient conditions on the utility functions for the pricing kernel to exist (Appendix 3.C). But we object to the dismal theorem for two reasons.

First, we think that the result is implied by using an incorrectly specified model. A key ingredient in Weitzman's model is the power utility function. This popular utility function is characterized by constant relative risk aversion (CRRA). The assumption of CRRA (hence RRA(0) $>$ 0) is not appropriate when dealing with extremely low levels of consumption, and it is exactly the behavior at these low consumption levels that leads to the dismal theorem. As we have demonstrated in Section 3.4, Weitzman's result is avoided when the economic model (utility function) is compatible with the statistical model (heavy tails). Utility functions with appropriate risk aversion for low levels of consumption are not subject to the dismal theorem.

Second, more effort can be made to know an input parameter that is 'hard to know', and we describe a (stylized) learning-and-agreement procedure for precisely this purpose in Sections 3.5.1 and 3.5.2. Although it is difficult to state upper and lower bounds for the 'VSL-like' input parameter, we can still obtain reasonable constraints on difficult-to-know parameters of interest indirectly. The economic model translates the parameter of interest into output variables with an easier interpretation (such as the optimal policies and the probability of catastrophe). Bounds on these output variables, together with the economic model, imply bounds on the parameter of interest.

3.6 Conclusions

Our strategy in this paper has been to first specify and analyze a stochastic economy-climate model using the popular power utility function. Section 3.3 demonstrates explicitly that power utility is fragile with respect to distributional assumptions. This is not unexpected. Weitzman (2009) summarizes this fragility and the non-existence of a robust solution in a 'dismal' theorem.

We agree with Weitzman's concerns about the validity of expected utility analysis in settings featuring catastrophic risks. We argue that one should indeed allow for heavy-tailed distributions when modeling catastrophic climate changes, but that, in contrast to Weitzman, heavy-tailed distributional assumptions are not *per se* irreconcilable with expected utility.

Based on general results regarding the relationship between the richness of the class of utility functions and the generality of the permitted distributional assumptions (Section 3.4 and Appendix 3.C), we then restrict ourselves to utility functions that are compatible with our distributional assumptions. In Section 3.4 we propose that on the domain that contains the typically observed consumption levels, the utility function behaves power-like (CRRA) as is popular in macroeconomics and finance, while on the remote domain containing extreme adverse consumption shocks, the utility function exhibits exponential-like (CARA) features as is popular in insurance. Thus we avoid the unacceptable conclusion that society should sacrifice an unlimited amount of consumption to reduce the probability of catastrophic climate change by even a small amount. After reaching agreement on the model parameters, the sensitivity analysis conducted in Section 3.5 shows that our completed model and the resulting optimal policies are quite robust and sensibly sensitive. With quasi exponential-like behavior of the utility function in near-catastrophe situations, extreme sensitivities that would otherwise be present using CRRA preferences are avoided.

Much of the analysis in our paper is not limited to extreme climate change. A similar analysis could apply in other policy making settings involving catastrophic risks, such as the development of new financial incentive schemes to mitigate the risk of extreme systemic failures and resulting financial economic crises, or policies concerning medical risks (pandemic flu and vaccination risks).

Let us finally admit four limitations of our paper, and indicate possible generalizations. First, from Section 3.4 onwards, we have focussed our attention on bounded utility functions, so as to avoid having to restrict distributional assumptions. In general, one could assume more structure on stochasticity (yet still allow for heavy tails) and broaden the constraints on utility. In

particular, unbounded utility (such as HARA with $0 < \alpha \leqslant 1$; see Chapter 4 is also permitted under additional assumptions on stochasticity. Second, for simplicity and clarity of presentation, we have restricted our analysis to only two periods. In principle, much of our analysis will remain intact when considering more than two periods. Third, to account for the fact that the policy maker has the double objective of maximizing current consumption, while also leaving a reasonable economy for the next policy maker, we have used scrap values in our analysis. We ignore, however, stochasticity in the scrap value function after the second period. The development of a numerically tractable economy-climate model with multi-period stochasticity *and* scrap values is left for future research. Finally, the equations making up our stochastic economy-climate model are of a simple and stylized nature, and each one of them, including the specification of stochasticity, leaves room for generalizations and extensions.

3.A Kuhn-Tucker conditions under positive investment

Consider the economy-climate model of Section 3.2 in the two-period set-up of Section 3.3. Let U be a general well-behaved utility function and let $S^{(1)}$ and $S^{(2)}$ be general well-behaved scrap value functions. At the beginning of period 1 our welfare function, conditional on (C_0, μ_0, ϵ_1), is

$$W = L_1 U(C_1/L_1) + \nu_1 S^{(1)}(K_2) - \nu_2 S^{(2)}(M_2).$$

We have four constraints: $C_1 \geqslant 0$, $I_1 \geqslant 0$, $\mu_1 \geqslant 0$, and $\mu_1 \leqslant 1$, but only two of these can be binding as we shall see. Hence, we define the Lagrangian $\mathcal{L} = \mathcal{L}(C_1, \mu_1)$ as

$$\mathcal{L} = L_1 U(C_1/L_1) + \nu_1 S^{(1)}(K_2) - \nu_2 S^{(2)}(M_2) + \kappa_1 I_1 + \kappa_2(1 - \mu_1),$$

and we find

$$\frac{\partial \mathcal{L}}{\partial C_1} = U'(C_1/L_1) - (\nu_1 g_1 + \kappa_1)$$

and

$$\frac{\partial \mathcal{L}}{\partial \mu_1} = \left(-(\nu_1 g_1 + \kappa_1)\psi_1 \theta \mu_1^{\theta-1} d_1 + \nu_2 g_2 \sigma_1\right) Y_1 - \kappa_2,$$

where

$$g_1 = g_1(C_1, \mu_1) = \frac{\partial S^{(1)}(K_2)}{\partial K_2}, \qquad g_2 = g_2(\mu_1) = \frac{\partial S^{(2)}(M_2)}{\partial M_2}.$$

This leads to the Kuhn-Tucker conditions:

$$\kappa_1 = U'(C_1/L_1) - \nu_1 g_1 \geqslant 0,$$
$$I_1 = (1 - \psi_1 \mu_1^{\theta}) d_1 Y_1 - C_1 \geqslant 0,$$

and

$$\kappa_2 = \left(-U'(C_1/L_1)\psi_1 \theta \mu_1^{\theta-1} d_1 + \nu_2 g_2 \sigma_1\right) Y_1 \geqslant 0,$$
$$\mu_1 \leqslant 1,$$

together with the slackness conditions $\kappa_1 I_1 = 0$ and $\kappa_2(1 - \mu_1) = 0$.

Under the assumption that $I_1 > 0$ we have $\kappa_1 = 0$ and we distinguish between two cases, as follows.

Case (1): $\kappa_2 > 0$. We have $\mu_1 = 1$ and $g_2 = g_2(1)$, and we solve two equations in two unknowns:

$$U'(C_1/L_1) = \nu_1 g_1, \qquad g_1 = g_1(C_1, 1),$$

under the restrictions:

$$\frac{C_1}{(1 - \psi_1) Y_1} \leqslant d_1 < \frac{\nu_2 g_2 \sigma_1}{\nu_1 g_1 \psi_1 \theta}.$$

Case (2): $\kappa_2 = 0$. We solve four equations in four unknowns:

$$U'(C_1/L_1) = \nu_1 g_1, \qquad \mu_1^{\theta-1} d_1 = \frac{\nu_2 g_2 \sigma_1}{\nu_1 g_1 \psi_1 \theta},$$

$$g_1 = g_1(C_1, \mu_1), \qquad g_2 = g_2(\mu_1),$$

under the restrictions:

$$C_1 \leqslant (1 - \psi_1 \mu_1^{\theta}) d_1 Y_1, \qquad \mu_1 \leqslant 1.$$

The following two points are worth noting. First, we see that the restrictions $\mu_1 \geqslant 0$ and $C_1 \geqslant 0$ are automatically satisfied, so that they do not need to be imposed. Second, we see that $U'(C_1/L_1) = \nu_1 g_1$ in both cases. This fact will be used in Appendix 3.B.

3.B Proof of Proposition 3.1

We shall prove the proposition both for the linear scrap and the non-linear scrap case. In both cases the inequality constraints (3.2) are imposed. Since

$$d_1 Y_1 = B_1 e^{\tau \epsilon_1}, \qquad B_1 = \frac{e^{-\tau^2/2} Y_1}{1 + \xi H_1^2},$$

we obtain

$$C_1^* \leqslant C_1^* + I_1^* = (1 - \omega_1^*) d_1 Y_1 \leqslant B_1 e^{\tau \epsilon_1}, \tag{3.12}$$

$$I_1^* \leqslant C_1^* + I_1^* \leqslant B_1 e^{\tau \epsilon_1},$$

$$(1 - \delta) K_1 \leqslant K_2^* \leqslant (1 - \delta) K_1 + B_1 e^{\tau \epsilon_1},$$

and

$$M_2^* \leqslant (1 - \phi) M_1 + \sigma_1 Y_1.$$

We distinguish between three cases.

Linear scrap under normality. Linear scrap implies that $S^{(1)}(K_2) = K_2$ and $S^{(2)}(M_2) = M_2$. Since $\mathrm{E}(e^{\tau\epsilon_1})$ exists under normality, it follows that C_1^*, I_1^*, K_2^*, and M_2^* all have finite expectations, and therefore that $\mathrm{E}(W^*)$ exists if and only $\mathrm{E}(1/C_1^*)$ exists. For notational convenience we do not distinguish between the random variable ϵ_1 and its realization. With this slight abuse of notation, we write

$$\begin{aligned}\mathrm{E}(1/C_1^*) &= \int_{-\infty}^{\infty} (1/C_1^*)\, dF(\epsilon_1) = \int_{I_1^*=0} (1/C_1^*)\, dF(\epsilon_1) + \int_{I_1^*>0} (1/C_1^*)\, dF(\epsilon_1) \\ &= (1/B_1) \int_{I_1^*=0} \frac{e^{-\tau\epsilon_1}}{1-\omega_1^*}\, dF(\epsilon_1) + \int_{I_1^*>0} (1/C_1^*)\, dF(\epsilon_1) \\ &\leqslant \frac{1}{(1-\psi_1)B_1}\, \mathrm{E}(e^{-\tau\epsilon_1}) + \int_{I_1^*>0} (1/C_1^*)\, dF(\epsilon_1).\end{aligned}$$

Since $\mathrm{E}(e^{-\tau\epsilon_1})$ is finite, it suffices to show that $\int_{I_1^*>0}(1/C_1^*)\, dF(\epsilon_1)$ is finite. Now, it follows from Appendix 3.A that, under the assumption that $I_1^* > 0$, $U'(C_1^*/L_1) = L_1^2/{C_1^*}^2 = \nu_1 g_1^* = \nu_1$, because $g_1^* = 1$. Hence,

$$\int_{I_1^*>0} (1/C_1^*)\, dF(\epsilon_1) = \frac{\nu_1^{1/2}}{L_1} \Pr(I_1^* > 0) \leqslant \frac{\nu_1^{1/2}}{L_1} < \infty.$$

Nonlinear scrap under normality. Nonlinear scrap implies that

$$S^{(1)}(K_2) = -\frac{K_0}{p}\left(\frac{K_2}{K_0}\right)^{-p}, \qquad S^{(2)}(M_2) = \frac{M_0}{q}\left(\frac{M_2}{M_0}\right)^{q}$$

where $p > 0$ and $q > 1$. Since

$$(K_2^*)^{-p} \leqslant ((1-\delta)K_1)^{-p}$$

and

$$(M_2^*)^{q} \leqslant ((1-\phi)M_1 + \sigma_1 Y_1)^{q},$$

we see that $\mathrm{E}(W^*)$ exists if and only $\mathrm{E}(1/C_1^*)$ exists. As in the linear scrap case, it suffices to show that $\int_{I_1^*>0}(1/C_1^*)\,dF(\epsilon_1)$ is finite. Since

$$g_1 = g_1(K_2) = \frac{\partial S^{(1)}(K_2)}{\partial K_2} = \left(\frac{K_0}{K_2}\right)^{p+1},$$

it follows from Appendix 3.A that, under the assumption that $I_1^* > 0$,

$$U'(C_1^*/L_1) = L_1^2/{C_1^*}^2 = \nu_1 g_1^* = \nu_1 \left(\frac{K_0}{K_2^*}\right)^{p+1} \leqslant \nu_1 \left(\frac{K_0}{(1-\delta)K_1}\right)^{p+1},$$

and hence that

$$\int_{I_1^*>0}(1/C_1^*)\,dF(\epsilon_1) \leqslant \frac{\nu_1^{1/2}}{L_1}\left(\frac{K_0}{(1-\delta)K_1}\right)^{(p+1)/2} < \infty.$$

Student distribution. From (3.12) we have $1/C_1^* \geqslant e^{-\tau\epsilon_1}/B_1$. Under a Student distribution, the right-hand side has no finite expectation, and hence the left-hand side has no finite expectation either. In the non-linear scrap case, this is sufficient to prove the non-existence of $\mathrm{E}(W^*)$ because $S^{(1)}(K_2^*)$ and $S^{(2)}(M_2^*)$ are both bounded. In the linear scrap case, M_2^* is bounded, but K_2^* is not. Now, since

$$C_1^* \leqslant B_1 e^{\tau\epsilon_1}, \qquad K_2^* \leqslant (1-\delta)K_1 + B_1 e^{\tau\epsilon_1},$$

we obtain

$$L_1(1 - L_1/C_1^*) + \nu_1 K_2^* \leqslant L_1 - (L_1^2/B_1)\,e^{-\tau\epsilon_1} + \nu_1(1-\delta)K_1 + \nu_1 B_1 e^{\tau\epsilon_1} \equiv G(\epsilon_1).$$

Since G is monotonically increasing from $-\infty$ to $+\infty$, there exists a unique ϵ_1^* defined by $G(\epsilon_1^*) = 0$. Hence, $G(\epsilon_1) \leqslant 0$ for all $\epsilon_1 \leqslant \epsilon_1^*$ and

$$\begin{aligned}\mathrm{E}\,|(L_1(1 - L_1/C_1^*) + \nu_1 K_2^*)| &\geqslant \int_{\epsilon_1 \leqslant \epsilon_1^*} |G(\epsilon_1)|\, dF(\epsilon_1) \\ &\geqslant -L_1 - \nu_1(1-\delta)K_1 + (L_1^2/B_1) \int_{\epsilon_1 \leqslant \epsilon_1^*} e^{-\tau\epsilon_1}\, dF(\epsilon_1) - \nu_1 B_1 e^{\tau\epsilon_1^*} = \infty.\end{aligned}$$

3.C Expected utility and tail uncertainty

We now formulate our decision under uncertainty problem in Savage (1954) style, independent of the specific model considered in this paper, so that the results obtained below are generally applicable. We fix a set $\mathcal{S}$ of states of nature and we let $\mathcal{A}$ denote a σ-algebra of subsets of $\mathcal{S}$. One state is the true state. We also fix a set $\mathcal{C}$ of consequences (outcomes, consumption) endowed with a σ-algebra $\mathcal{F}$. Since we are only interested in monetary outcomes, we may take $\mathcal{C} = \mathbb{R}_+$. A decision alternative (policy bundle) X is a measurable mapping from $\mathcal{S}$ to $\mathcal{C}$, so that $X^{-1}(A) \in \mathcal{A}$ for all events $A \in \mathcal{F}$. We assume that the class of all decision alternatives $\mathcal{X}$ is endowed with a preference order $\succeq$.

Definition 3.1. We say that expected utility (EU) holds if there exists a measurable and strictly increasing function $U : \mathcal{C} \to \mathbb{R}$ on the space of consequences, referred to as the utility function, and a probability measure $\mathbb{P}$ on $\mathcal{A}$, such that the preference order $\succeq$ on $\mathcal{X}$ is represented by a functional V of the form $X \mapsto \int_{\mathcal{S}} U(X(s))\, d\mathbb{P} = V(X)$. Thus, the decision alternative $X \in \mathcal{X}$ is preferred to the decision alternative $Y \in \mathcal{X}$ if, and only if, $V(X) \geqslant V(Y)$.

In the Von Neumann and Morgenstern (1944) framework, utility U is subjective, whereas the probability measure $\mathbb{P}$ associated with $\mathcal{A}$ is objective and known beforehand (decision under risk). In the more general framework of Savage (1954) adopted here, the probability measure itself can be, but need not be, subjective (decision under uncertainty).

DEFINITION 3.2. We say that a risk $\epsilon : \mathcal{S} \to \mathbb{R}$ is heavy-tailed to the left (right) under $\mathbb{P}$ if its moment-generating function is infinite: $\mathrm{E}\,(e^{\gamma\epsilon}) = \infty$ for any $\gamma < 0$ ($\gamma > 0$).

Examples of heavy-tailed risks are the Student, lognormal, and Pareto distributions. Heavy-tailed risks provide appropriate mathematical models for low-probability high-impact events, such as environmental catastrophes.

PROPOSITION 3.2. If EU is to discriminate univocally among all possible alternative outcome distributions, the utility function must be bounded.

Proof: See Menger (1934, p. 468) in the context of St. Petersburg-type lotteries, and also Arrow (1974) and Gilboa (2009, pp. 108-109). Menger (implicitly) assumes boundedness from below and demonstrates that boundedness from above should hold, and it is straightforward to generalize his result to an a priori unrestricted setting. ‖

Proposition 3.2 states that the EU functional is finite for all outcome distributions if, and only if, the utility function is bounded. Moreover, the axiomatization of EU is valid for all outcome distributions if, and only if, the utility function is bounded. The implications are non-trivial: boundedness of the utility function must hold not just in exotic situations but also in more familiar and economically relevant settings involving high levels of uncertainty. (See Moscadelli (2004) regarding operational risk.)

In what follows we do not require the utility function to be bounded. We simply assume that the class of feasible outcome distributions is restricted (though the restriction may be void) in such a way that the utility function permits discrimination among them. Recall from (3.5) that $\mathrm{RRA}(x) = -xU''(x)/U'(x)$ and $\mathrm{ARA}(x) = -U''(x)/U'(x)$, and let

$$\alpha^* = \inf_{x>0} \mathrm{RRA}(x), \qquad \beta^* = \sup_{x>0} \mathrm{ARA}(x).$$

Now consider a representative agent with time-additive EU preferences and time-preference parameter $\rho > 0$. We normalize (without loss of generality)

the agent's consumption C by setting $C_0 = 1$, and we define the pricing kernel (intertemporal marginal rate of substitution) as

$$P(C_1^*) = \frac{U'(C_1^*)}{(1+\rho)U'(1)}, \tag{3.13}$$

where C_1^* is optimal consumption at $t = 1$. Consumption C_1 is commonly restricted to a budget-feasible consumption set which is subject to uncertainty (ϵ_1). We assume that the budget restriction takes the general form

$$C_1^*(\epsilon_1) \leqslant B\exp(A\epsilon_1), \qquad B, A > 0, \tag{3.14}$$

which need not be best-possible. (In our economy-climate model of Section 3.2 as well as in the two-period setup of Section 3.3, $B = B_1$ and $A = \tau$.) The expectation $\mathrm{E}(P)$ represents the amount of consumption in period 0 that the representative agent is willing to give up in order to obtain one additional certain unit of consumption in period 1.

The following result states that the expectation of the pricing kernel is finite for all outcome distributions whenever the concavity index (Arrow-Pratt index, index of absolute risk aversion) $\mathrm{ARA}(x)$ is bounded.

Proposition 3.3. Assume that EU holds and that the budget feasibility restriction (3.14) applies.

(a) If $\alpha^* > 0$ and ϵ_1 is heavy-tailed to the left under $\mathbb{P}$, then $\mathrm{E}(P) = \infty$;

(b) If $\beta^* < \infty$ and $\alpha^* = 0$, then $\mathrm{E}(P) < \infty$ for any ϵ_1.

Proof: Let $\alpha^* > 0$. The EU maximizer is then more risk-averse in the sense of Arrow-Pratt than an agent with power (CRRA) utility of index α^*. It follows from (3.13) that

$$\frac{P'(C_1^*)}{P(C_1^*)} = \frac{U''(C_1^*)}{U'(C_1^*)} = -\mathrm{ARA}(C_1^*).$$

Since $\mathrm{ARA}(x) = \mathrm{RRA}(x)/x \geqslant \alpha^*/x$, we then have

$$\begin{aligned}\mathrm{E}(P) &= \frac{1}{1+\rho}\,\mathrm{E}\exp\left(-\int_{C_1^*}^1 d\log P(x)\right) = \frac{1}{1+\rho}\,\mathrm{E}\exp\left(\int_{C_1^*}^1 \mathrm{ARA}(x)\,dx\right)\\ &\geqslant \frac{1}{1+\rho}\int_{C_1^*\leqslant 1}\exp\left(\int_{C_1^*}^1(\alpha^*/x)\,dx\right)\,dF(\epsilon_1)\\ &= \frac{1}{1+\rho}\int_{C_1^*\leqslant 1}(C_1^*)^{-\alpha^*}\,dF(\epsilon_1) \geqslant \frac{B_1^{-\alpha^*}}{1+\rho}\int_{C_1^*\leqslant 1} e^{-\tau\alpha^*\epsilon_1}\,dF(\epsilon_1) = \infty,\end{aligned}$$

using (3.14) and the fact that ϵ_1 is heavy-tailed to the left. This proves part (a). Intuitively, if agent 1 is more risk-averse in the sense of Arrow-Pratt than agent 2, and if it is optimal to postpone all consumption for agent 2, then this will also be optimal for agent 1.

Next let $\alpha^* = 0$ and $\beta^* < \infty$. The EU maximizer is then less risk-averse in the sense of Arrow-Pratt than an agent with exponential (CARA) utility of index β^*. Since $\alpha^* = 0$, we have $0 \leqslant \mathrm{ARA}(x) \leqslant \beta^*$ and hence

$$\begin{aligned}\mathrm{E}(P) &= \int_{C_1^*\leqslant 1} P\,dF(\epsilon_1) + \int_{C_1^*>1} P\,dF(\epsilon_1)\\ &\leqslant \frac{1}{1+\rho}\int_{C_1^*\leqslant 1}\exp\left(\int_{C_1^*}^1 \beta^*\,dx\right)\,dF(\epsilon_1)\\ &\quad + \frac{1}{1+\rho}\int_{C_1^*>1}\exp\left(-\int_1^{C_1^*}\mathrm{ARA}(x)\,dx\right)\,dF(\epsilon_1)\\ &\leqslant \frac{e^{\beta^*}\Pr(C_1^*\leqslant 1)+\Pr(C_1^*>1)}{1+\rho} < \infty. \quad \|\end{aligned}$$

If the EU maximizer has decreasing absolute risk aversion and increasing relative risk aversion, as is commonly assumed, a complete and elegant characterization of boundedness of the expected pricing kernel can be obtained, as follows.

Proposition 3.4. Assume that EU holds and that the budget feasibility

restriction (3.14) applies. Assume furthermore that $\mathrm{RRA}(x)$ exists and is non-negative and non-decreasing for all $x \geqslant 0$ and that $\mathrm{ARA}(x)$ is non-increasing for all $x > 0$. Then, $\mathrm{E}(P) < \infty$ for any ϵ_1 if and only if $\int_0^\gamma \mathrm{ARA}(x)\,dx < \infty$ for some $\gamma > 0$.

Proof: To prove the 'only if' part, we assume that $\int_0^\gamma \mathrm{ARA}(x)\,dx$ is infinite for every $\gamma > 0$, and then show that there exist $(\mathcal{S}, \mathcal{A}, \mathbb{P})$ and ϵ_1 defined on it such that $\mathrm{E}(P) = \infty$. We note that $\beta^* = \infty$. Define a function $g : (0, 1] \to [1, \infty)$ by

$$g(y) = \exp\left(\int_y^1 \mathrm{ARA}(x)\,dx\right).$$

Then,

$$\mathrm{E}(P) \geqslant \frac{1}{1+\rho} \int_{C_1^* \leqslant 1} g(\min(C_1^*, 1))\,dF(\epsilon_1).$$

Recall from (3.14) that $C_1^* \leqslant B_1 e^{\tau\epsilon_1}$, and let ϵ_1^* be such that $B_1 e^{\tau\epsilon_1^*} = 1$, so that $0 < B_1 e^{\tau\epsilon_1^*} \leqslant 1$ if and only if $\epsilon_1 \leqslant \epsilon_1^*$. Define $u : (-\infty, \infty) \to [0, \infty)$ by

$$u(\epsilon_1) = \begin{cases} g(B_1 e^{\tau\epsilon_1}) - 1 & \text{if } \epsilon_1 \leqslant \epsilon_1^*, \\ 0 & \text{if } \epsilon_1 > \epsilon_1^*. \end{cases}$$

Since $\mathrm{ARA}(1) > 0$, g is monotonically decreasing and we obtain

$$\begin{aligned} \int_{C_1^* \leqslant 1} g(\min(C_1^*, 1))\,dF(\epsilon_1) &\geqslant \int_{\epsilon_1 \leqslant \epsilon_1^*} g(B_1 e^{\tau\epsilon_1})\,dF(\epsilon_1) \\ &= \int_{\epsilon_1 \leqslant \epsilon_1^*} (u+1)\,dF(\epsilon_1) = \mathrm{E}(u) + \Pr(\epsilon_1 \leqslant \epsilon_1^*). \end{aligned}$$

Strict monotonicity of g implies its invertibility. Hence we can choose u to be any non-negative random variable whose expectation does not exist (for example, the absolute value of a Cauchy distribution), and then define ϵ_1 through $B_1 e^{\tau\epsilon_1} = g^{-1}(u+1)$. With such a choice of ϵ_1 we have $\mathrm{E}(P) = \infty$.

To prove the 'if'-part we assume that $\int_0^\gamma \mathrm{ARA}(x)\,dx$ is finite. This implies that $\int_0^1 \mathrm{ARA}(x)\,dx$ is finite, so that

$$\begin{aligned}\mathrm{E}(P) &= \frac{1}{1+\rho}\int_{C_1^*\leqslant 1}\exp\left(\int_{C_1^*}^1 \mathrm{ARA}(x)\,dx\right)\,dF(\epsilon_1)\\ &\quad+\frac{1}{1+\rho}\int_{C_1^*>1}\exp\left(-\int_1^{C_1^*}\mathrm{ARA}(x)\,dx\right)\,dF(\epsilon_1)\\ &\leqslant \frac{\Pr(C_1^*\leqslant 1)}{1+\rho}\exp\left(\int_0^1\mathrm{ARA}(x)\,dx\right)+\frac{\Pr(C_1^*>1)}{1+\rho}<\infty,\end{aligned}$$

using the fact that $\alpha^* = \mathrm{RRA}(0) = 0$. ‖

Notice that, when $\int_0^\gamma \mathrm{ARA}(x)\,dx = \infty$ for some $\gamma > 0$, both $\alpha^* > 0$ and $\alpha^* = 0$ can hold. If $\alpha^* > 0$ then we do not need the full force of Proposition 3.4; it is sufficient that ϵ_1 is heavy-tailed to the left. Then $\mathrm{E}(P) = \infty$ by Proposition 3.3(a). If $\alpha^* = 0$ then heavy-tailedness alone is not sufficient, but we can always find an ϵ_1 such that $\mathrm{E}(P) = \infty$. An example of an ARA satisfying $\int_0^\gamma \mathrm{ARA}(x)\,dx = \infty$ and $\alpha^* = 0$ is a function which behaves as $-1/(x\log x)$ for values of x close to 0 and in addition satisfies the conditions of the proposition.

When $\int_0^\gamma \mathrm{ARA}(x)\,dx = \infty$ then $\beta^* = \infty$. But when $\int_0^\gamma \mathrm{ARA}(x)\,dx < \infty$, both $\beta^* < \infty$ and $\beta^* = \infty$ can occur. For example, when $\mathrm{ARA}(x) = x^{-\delta}$ $(0 < \delta < 1)$ then $\beta^* = \infty$; but when $\mathrm{ARA}(x) = \beta$ $(0 \leqslant \beta < \infty)$ then β^* is finite. A sufficient condition for $\int_0^\gamma \mathrm{ARA}(x)\,dx < \infty$ to hold is that there exists $0 \leqslant \delta < 1$ such that $\limsup_{x\downarrow 0} x^\delta \mathrm{ARA}(x) < \infty$.

Chapter 4

Burr utility

Abstract: This chapter proposes the Burr utility function. Burr utility is a flexible two-parameter family that behaves approximately power-like remote from the origin, while exhibiting exponential-like features near the origin. It thus avoids the extreme behavior of the power family near the origin, and is particularly suited for heavy-tailed risk analysis. We show how to characterize Burr utility as a special case in the general class of utility functions with non-increasing and convex absolute risk aversion, and non-decreasing and concave relative risk aversion. We further show its connection to the Burr probability distribution. A related class of generalized exponential utility functions is also studied.

4.1 Introduction

In most decision theories, including expected utility and the most common non-expected utility theories, the utility function U is unique up to positive affine transformations, that is, U is a cardinal (or interval) scale. In searching for a suitable utility function, it is the curvature of the function that is of interest. Since the second derivative U'' is not invariant to positive affine transformations in U, we typically normalize the second derivative by dividing

by the first (de Finetti, 1952; Pratt, 1964; Yaari, 1969; Arrow, 1971). It gives the Arrow-Pratt measure of absolute risk aversion:

$$\mathrm{ARA}(x) = \frac{-d\log U'(x)}{dx} = \frac{-U''(x)}{U'(x)}. \tag{4.1}$$

This degree of curvature is also referred to as the concavity index, a name that is particularly proper in non-expected utility theories, where concavity of U and risk aversion can not be identified. It captures all information for cardinal scales.

Up to positive affine transformations, there is precisely one function with *constant* absolute risk aversion (CARA), namely the cumulative distribution function of the exponential distribution,

$$U(x) = 1 - e^{-x/\lambda} \qquad (\lambda > 0). \tag{4.2}$$

Equally important is relative risk aversion:

$$\mathrm{RRA}(x) = \frac{-d\log U'(x)}{d\log x} = \frac{-xU''(x)}{U'(x)}. \tag{4.3}$$

Again, there exists precisely one function with *constant* relative risk aversion (CRRA), namely the power function,

$$U(x) = \frac{x^{1-\alpha} - 1}{1-\alpha} \qquad (\alpha > 0). \tag{4.4}$$

Throughout, we consider only non-negative inputs ($x \geqslant 0$). Exponential utility is bounded from above and below, and satisfies $\mathrm{ARA}(x) = 1/\lambda$ and $\mathrm{RRA}(x) = x/\lambda$. Thus, it exhibits constant ARA and increasing RRA. Power utility is either unbounded from above ($0 < \alpha < 1$) or from below ($\alpha > 1$) or both ($\alpha = 1$), and satisfies $\mathrm{ARA}(x) = \alpha/x$ and $\mathrm{RRA}(x) = \alpha$. Thus, it exhibits decreasing ARA and constant RRA.

In both theory and applications, power and exponential utility—in this order—are the most commonly used parametric families of utility functions.

They may perform credibly, but only if a restricted range of x is considered. If we are interested in inputs x remote from 0, as is common in macroeconomics and finance, then power utility is often appropriate (Wakker, 2008, and references therein). If, on the other hand, we are interested in near-catastrophe cases (small x), as in the insurance literature, then exponential utility is often used (Gerber, 1979, Chapter 5), thus avoiding the extreme behavior of power utility near $x = 0$. (We return to this issue at the end of Section 4.3 and in Section 4.6.) But if we are interested in the whole non-negative range of inputs, then more flexible families can be required. For example, Rabin (2000, p. 1287) indicates that CRRA preferences should not be used when both large and small inputs are relevant.

In this chapter we propose the Burr function, a utility function which behaves approximately power-like for inputs remote from 0 and exhibits exponential-like features for inputs near 0. As we will explain below, Burr utility is particularly suited for heavy-tailed risk analysis. In Section 4.2 we provide a characterization of an important class of utility functions, namely those where ARA is non-increasing and convex, and RRA is non-decreasing and concave. This is class $\mathcal{U}$. We restrict our attention to members of class $\mathcal{U}$. Two subclasses suggest themselves. In Section 4.3 we study the HARA subclass of which the Burr function is a special case. Section 4.4 provides a further and novel rationale for the Burr function. In Section 4.5 we study another subclass of $\mathcal{U}$, leading to 'gexpo' utility, a generalization of exponential utility, which is of independent interest. Section 4.6 provides a comparison of four utility functions, including the Burr function. Section 4.7 concludes.

4.2 Characterization of the $\mathcal{U}$ class

Motivated by the fact that both exponential and power utility exhibit non-increasing ARA and non-decreasing RRA in the spirit of Arrow (1971), we shall consider the following class of functions.

Definition 4.1. Let $U(x)$ be defined for $x \geqslant 0$ such that $U'(x) > 0$ and $U''(x) < 0$ for $x > 0$. If $\mathrm{ARA}(x)$ is non-increasing and convex and if $\mathrm{RRA}(x)$

is non-decreasing and concave for all $x > 0$, then we say that the function U belongs to the class $\mathcal{U}$. If, in addition, $\mathrm{RRA}(x) \to 0$ as $x \to 0$, then we say that U belongs to the class $\mathcal{U}_0$.

The importance of the reduction from $\mathcal{U}$ to $\mathcal{U}_0$ lies in the fact that if $\mathrm{RRA}(0) > 0$ then the expected intertemporal marginal rate of substitution (or pricing kernel) does not exist (is infinite) in the presence of heavy-tailed risks (Chapter 3). Note that exponential utility belongs to $\mathcal{U}_0$, while power utility belongs to $\mathcal{U}$ but not to $\mathcal{U}_0$.

It will be useful to define absolute risk tolerance as

$$T(x) = \frac{-U'(x)}{U''(x)} = \frac{1}{\mathrm{ARA}(x)} = \frac{x}{\mathrm{RRA}(x)}. \tag{4.5}$$

We have $T(x) = \lambda$ for exponential utility and $T(x) = x/\alpha$ for power utility, and hence both functions exhibit linear absolute risk tolerance. Utility functions with linear absolute risk tolerance are said to display 'harmonic absolute risk aversion', and are particularly useful to derive analytical results (Gollier, 2001, Section 2.6). Defining

$$R_1(x) = \frac{xT'(x)}{T(x)}, \qquad R_2(x) = \frac{x^2T''(x)}{T(x)}, \tag{4.6}$$

we obtain the following result.

Proposition 4.1. Assuming that $T(x)$ is twice differentiable, the class $\mathcal{U}$ is characterized by the inequalities

$$0 \leqslant R_1(x) \leqslant 1, \qquad -2R_1(x)(1 - R_1(x)) \leqslant R_2(x) \leqslant 2R_1^2(x).$$

If, in addition, $T(x)/x \to \infty$ as $x \to 0$, then we obtain the class $\mathcal{U}_0$.

Proof. Differentiating $\mathrm{ARA}(x) = 1/T(x)$ and $\mathrm{RRA}(x) = x/T(x)$ twice with respect to x, we find

$$\mathrm{ARA}'(x) \leqslant 0 \iff R_1(x) \geqslant 0, \qquad \mathrm{RRA}'(x) \geqslant 0 \iff R_1(x) \leqslant 1,$$

$$\mathrm{ARA}''(x) \geqslant 0 \iff R_2(x) \leqslant 2R_1^2(x),$$

and

$$\mathrm{RRA}''(x) \leqslant 0 \iff R_2(x) \geqslant -2R_1(x)(1 - R_1(x)). \;\|$$

For power utility we have $R_1 \equiv 1$, for exponential utility we have $R_1 \equiv 0$, and these two utility functions are therefore corner cases in $\mathcal{U}$. For both power and exponential utility we have $R_2 \equiv 0$. The two cases $R_2 \equiv 0$ and $R_1 \equiv r$ $(0 \leqslant r \leqslant 1)$ thus suggest themselves as natural extensions to power and exponential utility, and we analyze these two cases in Sections 4.3 and 4.5, respectively.

4.3 The class $R_2 \equiv 0$: HARA utility

The class $R_2 \equiv 0$ is characterized by $T(x) = ax + b$, and hence contains all utility functions that display linear harmonic absolute risk aversion (HARA). Since in this case $R_1(x) = ax/(ax + b)$, we find that U belongs to $\mathcal{U}$ if and only if $a \geqslant 0$, $b \geqslant 0$, and $a + b > 0$. If $a \geqslant 0$ and $b > 0$, then U belongs to $\mathcal{U}_0$. From (4.1) and (4.5) we find that the HARA class is also characterized by the differential equation

$$d \log U'(x) + \frac{dx}{ax + b} = 0 \qquad (a \geqslant 0, b \geqslant 0, a + b > 0).$$

There are three cases. For $a = 0$ and $b > 0$ we obtain the CARA utility function (exponential), for $a > 0$ and $b = 0$ we obtain the CRRA utility function (power), and for $a > 0$ and $b > 0$ we obtain the utility function in two steps. We first solve $U'(x) = A(ax + b)^{-1/a}$, and then, letting $\alpha = 1/a$

and $\lambda = b/a$,

$$U(x) = \frac{(x+\lambda)^{1-\alpha} - 1}{1-\alpha} \qquad (\alpha > 0, \lambda > 0), \tag{4.7}$$

apart from positive affine transformations. We see that

$$\mathrm{ARA}(x) = \frac{\alpha}{x+\lambda}, \qquad \mathrm{RRA}(x) = \frac{\alpha x}{x+\lambda}.$$

Both RRA and ARA are bounded in this case. When $0 < \alpha \leqslant 1$ utility is bounded from below but unbounded from above; when $\alpha > 1$ utility is bounded. Marginal utility is bounded for every α, also at zero.

We conclude that the HARA class contains seven types of utility functions, as follows:

U does not belong to $\mathcal{U}_0$:

unbounded: $U(x) = \log x$,

bounded from below, but not from above: $U(x) = x^r \quad (0 < r < 1)$,

bounded from above, but not from below: $U(x) = 1 - x^{-k} \quad (k > 0)$;

U belongs to $\mathcal{U}_0$:

bounded from below, but not from above:

$U(x) = \log(1 + x/\lambda) \quad (\lambda > 0)$,

$U(x) = (1 + x/\lambda)^r - 1 \quad (\lambda > 0,\ 0 < r < 1)$,

bounded:

$U(x) = 1 - e^{-x/\lambda} \quad (\lambda > 0)$,

$U(x) = 1 - (1 + x/\lambda)^{-k} \quad (\lambda > 0,\ k > 0)$,

where we have normalized the functions—without loss of generality—such that if there is a lower bound, it is zero; and if there an upper bound, it is one. The last of these seven functions is the so-called Burr utility function. It is bounded, belongs to $\mathcal{U}_0$, and has a number of other attractive features; see Section 4.4 below.

All members of the HARA class belong to $\mathcal{U}$. If a member of the HARA class does not belong to $\mathcal{U}_0$, then it belongs to the power family (and vice versa). In that case RRA(0) > 0 by definition, so that ARA(0) $= \infty$. If a member of the HARA class does belong to $\mathcal{U}_0$, then RRA(0) $= 0$ and *always* ARA(0) $< \infty$. The extreme behavior of the power family near $x = 0$, where ARA is unbounded, can generate important problems when inputs are not bounded away from 0 and risks feature heavy tails; see Chapter 3 for a detailed analysis. Modifying the units of inputs (to $\tilde{x} = ax, a > 0$) does not affect the power family—an exclusive property of this family—but does of course not remedy these problems. Modifying the level of inputs (to $\tilde{x} = x + b$) could conceivably solve the problems, but does affect the power family.

We note that for all utility functions in the intersection of HARA and $\mathcal{U}_0$, a modification of the units of inputs can be nullified by adjusting the parameter λ. Also, within the HARA class, only the exponential family is invariant to a modification of the level of inputs.

The utility function (4.7) has received some attention (Harrison *et al.*, 2007). It is an appealing, seemingly more appropriate, alternative to the power family (4.4). The two parameters α and λ jointly characterize the utility function (4.7), but individually don't have a specific empirical meaning. The behavior of (4.7) is quite different when $\lambda < 0$ would be assumed, as is the case for Stone-Geary utility functions. The parameter λ then plays the role of subsistence level. With $\lambda > 0$, RRA(0) $= 0$ and RRA is increasing, while with $\lambda < 0$, RRA$(-\lambda) = \infty$ and, for $x > -\lambda$, RRA is decreasing.

An additional problem with the power family is the extreme behavior of its derivatives at $x = 0$. It implies that in a setting—not considered here—with both positive and negative inputs including $x = 0$, the loss aversion index of Köbberling and Wakker (2005), defined as the ratio of one-sided derivatives at $x = 0$, behaves improperly under power utility. We note that, by contrast, all utility functions in the intersection of HARA and $\mathcal{U}_0$ are smooth at $x = 0$, and allow for a generalization to a setting with both positive and negative inputs that induces proper behavior of this loss aversion index.

4.4 Burr utility

There is an interesting connection between the HARA class and the Burr distribution. The Burr cumulative distribution function (Burr, 1942; Burr and Cislak, 1968) is defined by

$$U(x) = 1 - (1 + (x/\lambda)^c)^{-k} \qquad (k > 0, \lambda > 0, c > 0). \tag{4.8}$$

This is a three-parameter family of distribution functions with the property that many of the known distribution functions are special or limiting cases. It is therefore an appropriate function to approximate an unknown distribution function. Absolute risk tolerance is given by

$$T(x) = \frac{\lambda(1 + (x/\lambda)^c)(x/\lambda)}{(ck + 1)(x/\lambda)^c + (1 - c)}.$$

One verifies that U belongs to $\mathcal{U}$ if and only if $c \leqslant 1$, and that U belongs to $\mathcal{U}_0$ if and only if $c = 1$. For $c = 1$ we obtain

$$U(x) = 1 - \left(\frac{\lambda}{x + \lambda}\right)^k \qquad (k > 0, \lambda > 0), \tag{4.9}$$

corresponding to $T(x) = (x+\lambda)/(k+1)$. We call this function the *Burr utility function* and we see that it is precisely the HARA utility function (4.7) when $\alpha > 1$. If we think of the Burr family of distribution functions as a family of utility functions, and require only that utility is concave and that $\mathrm{RRA}(0) = 0$, then we obtain the non-exponential bounded HARA utility function, that is, Burr utility.

Burr utility has several appealing features. It is bounded, and satisfies $\mathrm{RRA}(0) = 0$ and $\mathrm{ARA}(0) < \infty$, properties that are particularly relevant when considering risks with arbitrarily heavy tails (Chapter 3). It further has increasing RRA as is empirically justified for decision under risk (Friend and Blume, 1975; Binswanger, 1980; Holt and Laury, 2002; Post *et al.*, 2008).

Finally, it behaves approximately power-like for inputs remote from 0, corresponding to empirical evidence (Chiappori and Paiella, 2008). No other member of the HARA class satisfies this combination of features.

Burr utility, like any member of the HARA class, has a completely monotone first derivative with higher order derivatives of alternating sign. Its index of n-th order risk attitude (Denuit and Eeckhoudt, 2010) is given by

$$(-1)^{n+1}\frac{U^{(n)}(x)}{U'(x)} = [(1+k)(2+k)\cdots((n-1)+k)](x+\lambda)^{-n+1}, \qquad n \geqslant 2.$$

4.5 The class $R_1 \equiv r$: gexpo utility

The class $R_1 \equiv r$ $(0 \leqslant r \leqslant 1)$ is characterized by $T(x) = x^r/\beta$, and we see that $R_2(x) = -r(1-r)$. Clearly, U belongs to $\mathcal{U}$; it belongs to $\mathcal{U}_0$ if and only if $\beta > 0$ and $0 \leqslant r < 1$. To solve $U(x)$ from $T(x)$ we first obtain marginal utility $U'(x)$ from

$$d\log U'(x) + \beta x^{-r}\,dx = 0.$$

This yields $U'(x) = A\exp(-\beta x^{1-r}/(1-r))$, where A is an arbitrary positive constant. Let us reparameterize by letting $p = 1/(1-r)$, excluding henceforth the power family $(r = 1)$, and choose A such that $U'(x)$ is a proper density function. Then,

$$U'(x) = \frac{(p\beta)^p e^{-h(x)}}{\Gamma(p+1)}, \qquad h(x) = p\beta x^{1/p} \qquad (p \geqslant 1, \beta > 0), \tag{4.10}$$

is a special case of the three-parameter generalized gamma density, other special cases of which include the two-parameter gamma, the Weibull, and the lognormal densities (Stacy, 1962; Johnson *et al.*, 1995). This density, first proposed by Subbotin (1923), is sometimes called the 'exponential power' or the 'power exponential' density (Johnson *et al.*, 1995, pp. 195–198); we shall call it the 'gexpo' density, because it generalizes the exponential density from $p = 1$ to $p \geqslant 1$. From the density function U' we obtain the cumulative

distribution function U as

$$U(x) = 1 - \frac{(p\beta)^p x h^{-p}(x)\,\Gamma\,(p, h(x)\,)}{\Gamma(p)}, \tag{4.11}$$

where $\Gamma(p,h) = \int_h^\infty t^{p-1}e^{-t}\,dt$ denotes the incomplete gamma function, and $\Gamma(p) = \Gamma(p,0)$ is the (complete) gamma function; see Abramowitz and Stegun, 1964, Chapter 5. This expression can not be simplified unless p is a positive integer, in which case we obtain

$$\Gamma\,(p, h(x)\,) = \Gamma(p)e^{-h(x)} \sum_{k=0}^{p-1} \frac{h^k(x)}{k!},$$

and hence

$$U(x) = 1 - e^{-h(x)} \sum_{k=0}^{p-1} \frac{h^k(x)}{k!}. \tag{4.12}$$

This specializes to exponential utility (4.2) when $p = 1$, and to

$$U(x) = 1 - e^{-2\beta\sqrt{x}}(1 + 2\beta\sqrt{x}) \tag{4.13}$$

when $p = 2$. Like Burr utility, gexpo utility is bounded from above and below, belongs to $\mathcal{U}_0$ and has smooth derivatives at $x = 0$. But unless $p = 1$, gexpo utility still exhibits extreme behavior near the origin, with $\mathrm{ARA}(0) = \infty$.

4.6 Comparison of four utility functions

In this section we shall compare, mostly graphically, the behavior of four members of the $\mathcal{U}$ class of utility functions:

exponential:

$$U_1(x) = 1 - e^{-x/\lambda_1}, \qquad U_1'(x) = (1/\lambda_1)e^{-x/\lambda_1},$$

power:

$$U_2(x) = \frac{x^{1-\alpha} - 1}{1-\alpha}, \qquad U_2'(x) = x^{-\alpha},$$

Burr:

$$U_3(x) = 1 - \left(\frac{\lambda_2}{x+\lambda_2}\right)^k, \qquad U_3'(x) = \frac{k\lambda_2^k}{(x+\lambda_2)^{k+1}},$$

gexpo:

$$U_4(x) = 1 - e^{-h(x)} \sum_{k=0}^{p-1} \frac{h(x)^k}{k!}, \qquad U_4'(x) = \frac{(p\beta)^p e^{-h(x)}}{p!},$$

where $h(x) = p\beta x^{1/p}$, and $\lambda_1 > 0$, $\alpha > 0$, $k > 0$, $\lambda_2 > 0$, $\beta > 0$, and $p \geqslant 1$. The ARA and RRA functions in the four cases are given by

$$\mathrm{ARA}_1 = 1/\lambda_1, \quad \mathrm{ARA}_2 = \frac{\alpha}{x}, \quad \mathrm{ARA}_3 = \frac{k+1}{x+\lambda_2}, \quad \mathrm{ARA}_4 = \frac{\beta}{x^{(p-1)/p}},$$

and

$$\mathrm{RRA}_1 = x/\lambda_1, \quad \mathrm{RRA}_2 = \alpha, \quad \mathrm{RRA}_3 = \frac{(k+1)x}{x+\lambda_2}, \quad \mathrm{RRA}_4 = \beta x^{1/p}.$$

In order to compare the four utility functions, we determine a point x^* where we want the four functions to be 'close'. Without affecting the results, let us choose $x^* = 0.08$. By 'close' we mean that $\mathrm{RRA}(x^*)$ is the same for each of the four functions. If we choose $\alpha = 2$, $k = 1.5$, and $p = 2$, then this condition implies $\lambda_1 = 0.04$, $\lambda_2 = 0.02$, and $\beta = 5\sqrt{2}$.

In Figure 4.1 we graph the (scaled) marginal utility $g(x) = U'(x)/U'(x^*)$, in the left panel for $0 < x < 0.2$ and in the right panel zoomed in closer to the point $x^* = 0.08$. The four graphs do not intersect, and, because of the normalizations, they are tangent at $x = x^*$. For $x \neq x^*$ we have

$$g_{power}(x) > g_{Burr}(x) > g_{gexpo}(x) > g_{exp}(x),$$

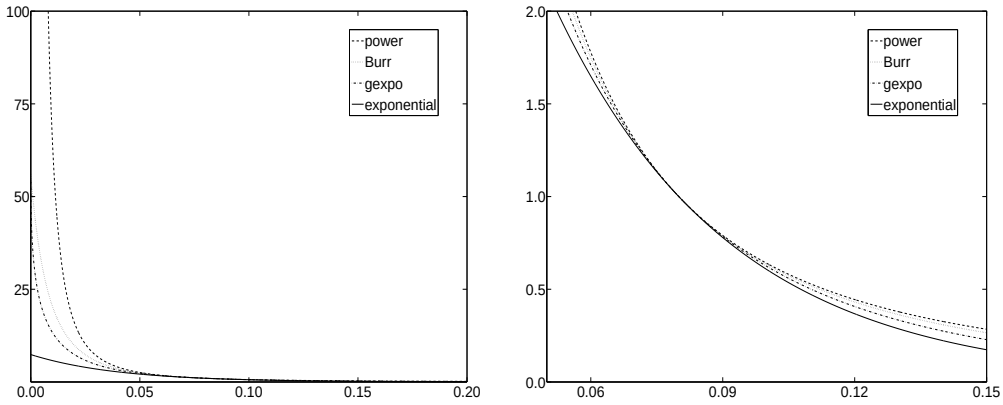


Figure 4.1: Marginal utility (scaled) for four utility functions

and, at $x = 0$,

$$g_{power}(0) = \infty, \quad g_{Burr}(0) = 55.9, \quad g_{gexpo}(0) = 54.6, \quad g_{exp}(0) = 7.4.$$

Marginal utility is bounded except for the power function.

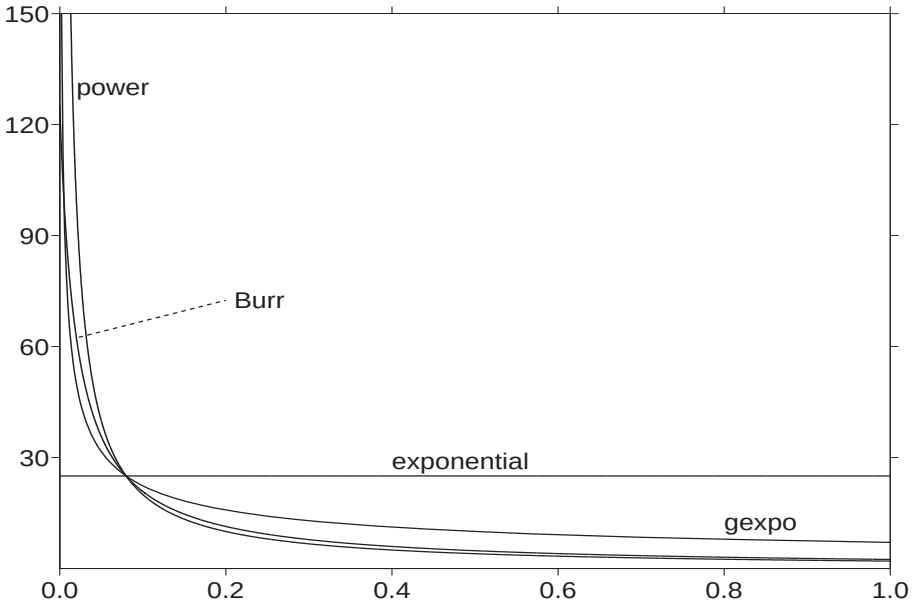


Figure 4.2: Absolute risk aversion for four utility functions

Absolute risk aversion ARA is graphed in Figure 4.2. We have, for $0 < x < 0.005$,

$$\mathrm{ARA}_{power}(x) > \mathrm{ARA}_{gexpo}(x) > \mathrm{ARA}_{Burr}(x) > \mathrm{ARA}_{exp}(x),$$

where $\mathrm{ARA}_{power}(0)$ and $\mathrm{ARA}_{gexpo}(0)$ are both infinite, and $\mathrm{ARA}_{Burr}(0) = 125$ and $\mathrm{ARA}_{exp}(0) = 25$; for $0.005 < x < x^*$ we have

$$\mathrm{ARA}_{power}(x) > \mathrm{ARA}_{Burr}(x) > \mathrm{ARA}_{gexpo}(x) > \mathrm{ARA}_{exp}(x);$$

and, for $x > x^*$,

$$\mathrm{ARA}_{power}(x) < \mathrm{ARA}_{Burr}(x) < \mathrm{ARA}_{gexpo}(x) < \mathrm{ARA}_{exp}(x).$$

We see that ARA (Burr) is very close to ARA (power) when $x > x^*$, but that this is no longer the case when x is close to zero, since ARA (Burr), just like ARA (exponential), remains finite while ARA (power), just like ARA (gexpo), goes to infinity.

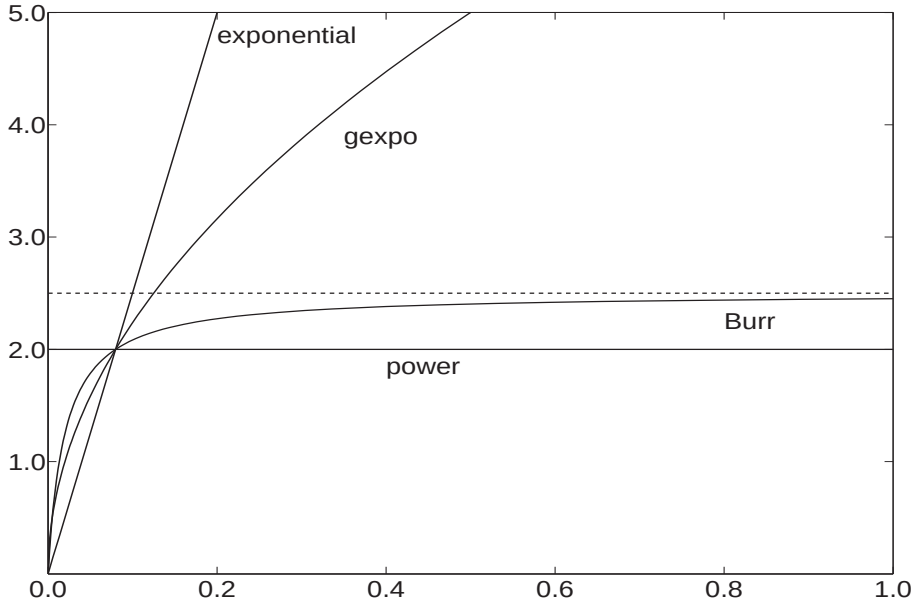


Figure 4.3: Relative risk aversion for four utility functions

In Figure 4.3 we graph relative risk aversion RRA. This shows that $\text{RRA}(0) = 0$ except for power utility, that both Burr and gexpo lie in-between power and exponential, and that, when x is large, gexpo behaves more like exponential and Burr more like power.

4.7 Conclusions

In this chapter we propose Burr utility, a flexible two-parameter family of utility functions. Burr utility enjoys a combination of appealing properties, not shared by any other member of the HARA class, nor by the gexpo class of utility functions introduced in this chapter. This combination of properties is particularly relevant in heavy-tailed risk analysis.

Chapter 5

Scrap value functions in dynamic decision problems

Abstract: We introduce an accurate, easily implementable, and fast algorithm to compute optimal decisions in discrete-time long-horizon welfare-maximizing problems. The algorithm is useful when interest is only in the decisions up to period T, where T is small, and especially in a stochastic framework. It relies on a flexible parametrization of the relationship between state variables and optimal total time-discounted welfare through scrap value functions. We demonstrate that this relationship depends on the boundedness, half-boundedness, or unboundedness of the utility function, and on whether a state variable increases or decreases welfare. We propose functional forms for this relationship for general classes of utility functions and explain how the parameters can be identified.

5.1 Introduction

Let us consider a Ramsey-type growth model, presented as a dynamic optimization problem. We denote by U a non-decreasing, concave utility function,

and we define welfare W as

$$W = \sum_{t=0}^{\infty} \beta^t U(x_t), \tag{5.1}$$

where β is the time-discount factor and $x = \{x_t\}_{t=0}^{\infty}$ is a sequence of (univariate) control variables. A typical decision maker will maximize W subject to restrictions involving state and control variables. The latter include not only x but also other (possibly multivariate) control variables $y = \{y_t\}_{t=0}^{\infty}$. Optimal welfare subject to the restrictions is obtained at (x^*, y^*).

In practice, with the exception of some simple specific cases, it is computationally not feasible to solve the infinite-horizon problem. One typically solves a finite-horizon problem, say over T periods. When T is large, for example 60 periods, then the long horizon may well approximate the infinite horizon, but when T is small, for example 2 periods, then the short horizon will not provide a good approximation.

Our interest is in the decisions in the short-horizon problem, for two reasons. Solving the full long-horizon problem may computationally be too burdensome, especially in a stochastic context. But even if it were not burdensome, decision makers typically have a short horizon, if only because their term of office is short.

Supposing then that the decision maker has a finite T-period horizon, we write welfare (5.1) as

$$W = \sum_{t=0}^{T-1} \beta^t U(x_t) + \beta^T S_T, \qquad S_T = \sum_{t=0}^{\infty} \beta^t U(x_{T+t}).$$

For every $T \geqslant 0$ we now define the *scrap value* as

$$S_T^* = \sum_{t=0}^{\infty} \beta^t U(x_{T+t}^*), \tag{5.2}$$

which depends on the optimal path $(x_T^*, x_{T+1}^*, \dots)$, but not on y^* or on $(x_0^*, \dots, x_{T-1}^*)$. If we would know the optimal path $\{x_t^*\}$ for $t \geqslant T$, then

maximizing W over all x and y can be achieved by maximizing

$$\overline{W} = \sum_{t=0}^{T-1} \beta^t U(x_t) + \beta^T S_T^* \tag{5.3}$$

over $x_0, \ldots, x_{T-1}$ and $y_0, \ldots, y_{T-1}$. This simple observation is our starting point. The scrap value S_T^* depends on the state variables at time T, and we shall call this functional relationship the *scrap value function*. In fact, this is precisely the *value function* in dynamic programming, and the recursion

$$S_0^* = \sum_{t=0}^{T-1} \beta^t U(x_t^*) + \beta^T S_T^*$$

is known as the Bellman equation.

Only in very special cases do we know the functional form of the scrap value S_T^* as a function of the state variables. Hence we need to approximate it. The purpose of this chapter is to explore the relationship between the form of the utility function and the form of the scrap value function, to propose simple flexible forms for the scrap value functions, and to discuss how the parameters in these flexible forms can be estimated or calibrated. Throughout we shall distinguish between state variables that increase welfare (G for 'good'), say capital, and state variables that decrease welfare (B for 'bad'), say pollution. We shall see that the specification of the scrap value function depends on whether a state variable is good or bad in a nontrivial manner.

This chapter can be viewed in two ways. One can see it as providing a solution to a *long-horizon* (stochastic) dynamic decision problem by considering a short horizon and treating the remainder (the scrap value function) appropriately. But one can also see it as providing a solution to a *short-horizon* problem where the decision maker has two objectives: to maximize welfare over a short horizon and to leave a 'reasonable' state for the next decision maker. The two views are conceptually different but mathematically equivalent.

Most studies of dynamic optimal decision problems in climate-economic

models consider an infinite horizon. For numerical reasons, the model is then approximated based on methods introduced by Barr and Manne (1967) and summarized in Lau et al. (2002). That is, a terminal constraint is imposed on investment such that investment in the final period is sufficient to cover growth plus depreciation (Böhringer et al., 2007). Alternatively, a constraint is added such that investment growth equals output growth in the final period (Doroodian and Boyd, 2003). In either case, this implies a linear scrap value function (Böhringer et al., 2007, p. 697; Lau et al., 2002). In fact, scrap value functions in environmental economics are always assumed to be linear, primarily for simplicity. We show that the linear scrap value function is not reconcilable with any utility function, not even when the utility function is unbounded. Such a simplifying assumption may be harmless when T is large, but not when T is small. Our method of parametrization allows for non-linearities while retaining simplicity. To avoid overparametrization, which could in theory further improve accuracy as in any parametric method, the scrap value functions we propose are parsimonious.

There is a substantial literature on the theory of approximate dynamic programming. Without attempting to review this literature, we mention that projection methods, described in Judd (1992) and reviewed in Christiano and Fisher (2000), have become standard tools for solving dynamic models in economics. More recently, Lau et al. (2002) proposed a method to improve the approximation to the infinite-horizon problem through a complementarity formulation. The procedure proposed in this chapter avoids the sequential optimization that is required for their method (p. 586, footnote 5). Finally, Dorofeenko et al. (2010) introduced an efficient algorithm for solving stochastic dynamic models. However, their algorithm assumes proximity to the steady state, which is not reasonable for small T.

The plan of the remainder of this chapter is as follows. In Section 5.2 we explore the relationship between utility function and scrap value function in more detail. In Sections 5.3 and 5.4 we discuss our solution method, given the choice of scrap value function, first in a deterministic framework and then in a stochastic framework. Different utility functions require different specifications of the scrap value functions. In Sections 5.5–5.8, we discuss unbounded,

partially bounded, and bounded utility functions in detail. Section 5.9 concludes.

5.2 Scrap value and utility functions

Since we do not know the scrap value S_T^* in (5.3), we need to approximate it. Some authors take T to be large but finite, and add a terminal condition to the model. For example, the 60-period (600 years) DICE model in Nordhaus (2008) includes a terminal condition, namely that at least 2% of the capital stock at the beginning of period T should be invested annually during period T; see also Doroodian and Boyd (2003, Section 3.6) and Leach (2009, Section 3). In fact, this terminal condition is equivalent to a linear scrap value function for capital (Eyckmans and Tulkens, 2003).

When T is large then $\beta^T S_T^* \approx 0$, and a poor approximation to the scrap value may not have large effects on the values of the optimal controls. But when T is small, we are forced to look for good approximations, thus emphasizing the fact that the decision maker has the double objective of maximizing time-discounted welfare over a finite number of periods T, while also leaving a reasonable economy for the next decision maker, based on the remaining stocks of something good ($G > 0$), say capital, and something bad ($B > 0$), say pollution.

With these two state variables we write the scrap value S_T^* as

$$S_T^* = S^*(G_T, B_T).$$

The simplest specification for the scrap value function S^* is the linear function

$$S_T^* = \nu_0 + \nu_1 G_T - \nu_2 B_T \qquad (\nu_1 > 0,\, \nu_2 > 0), \tag{5.4}$$

where the parameters ν_1 and ν_2 denote the scrap prices of capital and pollution at the beginning of period T. This scrap value function satisfies the minimum requirement that the decision maker will be happier if there is more

capital and less pollution at the end of the period. But the function has two problems. First, it is more common and more realistic to assume that the scrap value function is concave rather than linear. Second, the linear scrap value function is unbounded, both towards $+\infty$ and $-\infty$, although the utility function on which welfare is based may be bounded or half-bounded, and this mathematical property should carry over to the scrap value function.

The main purpose of this chapter is to propose scrap value functions that are appropriate for different types of utility functions. Throughout we shall retain separability, so that we can write

$$S_T^* = \nu_0 + \nu_1 S_g^*(G_T) - \nu_2 S_b^*(B_T) \qquad (\nu_1 > 0,\, \nu_2 > 0). \tag{5.5}$$

Our task then is to specify the functions S_g^* and S_b^*. Without loss of generality we normalize the scrap value functions such that

$$\left.\frac{\partial S_g^*(G)}{\partial G}\right|_{G=G_0} = \left.\frac{\partial S_b^*(B)}{\partial B}\right|_{B=B_0} = 1, \tag{5.6}$$

which guarantees that if we linearize $S^*(G_T, B_T)$ around (G_0, B_0), we find

$$S^*(G_T, B_T) \approx \text{constant} + \nu_1 G_T - \nu_2 B_T,$$

so that ν_1 and ν_2 can be interpreted as scrap prices, just as in the linear case.

Supported by findings in Wirl (1991), Eyckmans and Tulkens (2003), and Krawczyk (2005), we shall choose S_g^* to be increasing and concave, so that the more capital is left, the better, but at a decreasing rate.

More difficult is the choice of S_b^*. When utility is unbounded from below (as in Sections 5.5 and 5.6) we shall choose S_b^* to be increasing and convex, so that the more pollution is left, the worse, and at an increasing rate. This is supported by Montgomery (1972) who argued formally in favor of a monotonic and convex abatement cost function; see also Hoel and Karp (2002), Feng and Zhao (2006), and de Zeeuw (2008). But when utility is bounded from below (as in Sections 5.7 and 5.8), then S_b^* must be bounded from above, and

this causes a problem because there are no increasing convex functions that are bounded from above. In that case we shall assume that S_b^* is convex-concave, that is, convex for low values of B_T and concave for higher values. A rigorous analysis of the optimal management of a convex-concave resource was provided by Skiba (1978). Several of the papers in Dasgupta and Mäler (2003) contain economic analyses of ecosystems whose natural regeneration functions are convex-concave. The introduction of a convex-concave scrap value function is closely related to the idea of a threshold; see for example Ranjan and Shortle (2007) and Leandri (2009).

The specification of the scrap value functions S_g^* and S_b^* depends on the utility function. A utility function can be:

- unbounded (for example, $U(x) = \log x$);
- bounded from above but not from below ($U(x) = 1 - 1/x$);
- bounded from below but not from above ($U(x) = \sqrt{x}$, $U(x) = \log(x + 1)$); or
- bounded ($U(x) = 1 - e^{-x}$, $U(x) = x/(x + 1)$).

Two of these four types are unbounded from above, in which case infinite welfare can occur. This can happen, in principle, when $x \to \infty$ or $G \to \infty$, but also when $B \to 0$. Although we can understand that infinite pollution gives infinite misery, it is less credible that no pollution gives infinite welfare. Hence we make the following assumption throughout.

Assumption 1: *The scrap value function S_b^* is bounded from below.*

One of the features of this chapter is that we distinguish between two types of scrap value function, associated with a good and a bad stock, while most of the literature only allows one type (good). We shall see that the modeling of the scrap value function associated with pollution (the 'bad') is more complex than for capital (the 'good').

5.3 Deterministic framework

Although the introduction and treatment of scrap value functions is particularly important in a stochastic framework, it will be useful to consider the deterministic framework first. A well-known example of such a framework is Nordhaus' (2008) economic model of climate change (the DICE model). This model has three state variables: capital, CO2 concentration, and temperature; and two controls: per capita consumption (x) and the abatement fraction for CO2 (y). More capital is good, more CO2 is bad, and a higher temperature is also bad (at a global level). Nordhaus considers 60 periods (600 years), and he imposes a terminal condition in order to obtain the optimal paths for the two control variables.

If we are only interested in short-horizon decisions (say 2 periods, 20 years), we have two options. We can still calculate, from (5.1), the optimal paths over 60 periods, and then only consider the optimal paths over the first 2 periods. Alternatively, we can set $T = 2$ and introduce scrap value functions, one for capital (good) and one for CO2 concentration (bad). (In this model it appears to be unnecessary to have a scrap value for temperature.) The assumed utility function is $U(x) = 1 - 1/x$, and the functional forms of the two scrap value functions $S_g^*(G_T)$ and $S_b^*(B_T)$ depend on the functional form of the utility function, as will be made more precise in Section 5.6. Given the functional form of $S_g^*(G_T)$ and $S_b^*(B_T)$, the total scrap value function is then given by (5.5):

$$S^*(G_2, B_2) = \nu_0 + \nu_1 S_g^*(G_2) - \nu_2 S_b^*(B_2) \qquad (\nu_1 > 0,\, \nu_2 > 0).$$

We need to estimate (calibrate) ν_1, ν_2, and the parameters in S_g^* and S_b^*. This is done by solving the model for different values of (G_0, B_0). Each set of initial state values will generate solutions $(G_2, B_2, S^*(G_2, B_2))$, and from these generated data we estimate the scrap value parameters by some standard line fitting method—in our case nonlinear least squares. Once we have estimated the parameters in the scrap value function S^*, we can obtain the optimal solutions for the first two periods from (5.3) instead of (5.1).

In a deterministic framework there is no computational advantage in using scrap value functions instead of optimizing over all (in this case 60) periods. However, the formulation using scrap values provides a summary, highlighting the essence of the short-horizon decision problem by introducing the policy period of the current decision maker and the desired state at the end of this policy period. In an analysis of current versus future decisions, this allows us to investigate the sensitivity of current optimal controls with respect to small changes in the scrap value parameters.

5.4 Stochastic framework

The main advantage of using scrap value functions becomes apparent in a stochastic framework, for example a stochastic version of the DICE model, as proposed in Chapter 3 in the context of catastrophic risk in economy-climate models. Suppose again that $T = 2$ and that there is only one random shock, ϵ, with some known cumulative distribution function F. The decision maker has two decisions to make, one at the beginning of period 0 and one at the beginning of period 1. The shock is observed at the end of period 0, and its observed value is taken into account for the second decision. But for the first decision the decision maker does not know ϵ, and will therefore take into account its distribution but not its realization. Welfare is a function of the chosen sequence of controls, $x = \{x_t\}_{t=0}^{\infty}$, the initial values of the state variables (G_0, B_0), and the shock ϵ. Thus, conditional on (G_0, B_0), $W = W(x, \epsilon)$. At time 0 all that is known about ϵ is its distribution, and hence the decision maker maximizes expected welfare,

$$\mathrm{E}(W) = \int W(x, \epsilon)\, dF(\epsilon).$$

subject to the restrictions involving state and control variables.

Without introducing scrap value functions, the long-horizon problem may not be solvable or it may be computationally too burdensome. But with scrap value functions we can maximize expected welfare in four steps as follows.

First, determine the scrap value function $S^*(G_2, B_2)$ (both its structural form and the values of its parameters) by ignoring stochasticity, using the perturbation method described in the previous section. Then, write welfare as

$$\overline{W}(x_0, x_1, \epsilon) = U(x_0, \epsilon) + \beta U(x_1, \epsilon) + \beta^2 S^*(G_2, B_2),$$

and maximize $\overline{W}$ with respect to x_1 conditional on (x_0, ϵ). This gives x_1^* and concentrated welfare

$$\overline{W}_c(x_0, \epsilon) = \overline{W}(x_0, x_1^*, \epsilon).$$

Next, compute the expectation $\mathrm{E}(\overline{W}_c(x_0, \epsilon))$, if it exists. Finally, maximize this expectation with respect to x_0 using a standard grid-based approach.

The big computational advantage lies in the fact that for every iteration of the maximization and for each grid cell, we now only need to solve a short-horizon problem. Since the number of perturbations required to estimate the scrap value function is very much smaller than the number of grid cells required for the maximization, the computational burden of our problem is much reduced, and this reduction increases as T decreases.

5.5 Unbounded utility

The infinite-horizon problem with a given utility function implies the functional form of the scrap value function in the finite-horizon problem. But, in general, we can not derive this functional form. There are, however, certain properties of the utility function that carry over to the scrap value function, namely whether the function is bounded, half-bounded, or unbounded. By employing this simple fact we obtain much-improved approximations to the scrap value functions. In this and the next three sections we propose flexible functional forms for scrap value functions (both for 'good' and for 'bad' state variables) for each of four types of utility functions.

In the class of unbounded utility functions, the function $U(x) = \log(x)$ is probably the best-known example. It is unbounded from below and from above, and the corresponding scrap value functions should therefore also be

unbounded. Writing the scrap value function as

$$S_T^* = \nu_0 + \nu_1 S_g^*(G_T) - \nu_2 S_b^*(B_T),$$

the simplest option is the linear scrap value function (5.4) where $S_g^*(G_T) = G_T$ and $S_b^*(B_T) = B_T$. Although the linear scrap value function seems to be unbounded, in fact it is not, because $G_T > 0$ and $B_T > 0$, so that both S_g^* and S_b^* are only half-bounded.

A suitable alternative would be a scrap value function in the power class, namely

$$S_T^* = \nu_0 + \zeta_1 \log(G_T) - \zeta_2 B_T^q \qquad (\zeta_1 > 0,\, \zeta_2 > 0,\, q \geqslant 1).$$

Choosing ζ_1 and ζ_2 according to the normalization (5.6), we obtain

$$S_g^*(G_T) = G_0 \log(G_T), \qquad S_b^*(B_T) = \frac{B_0}{q}\left(\frac{B_T}{B_0}\right)^q, \tag{5.7}$$

where $q \geqslant 1$. Notice that S_g^* is increasing and strictly concave on $(0, \infty)$, and that S_b^* is increasing and convex on $(0, \infty)$ in accordance to the arguments in Section 5.2 and Assumption 1.

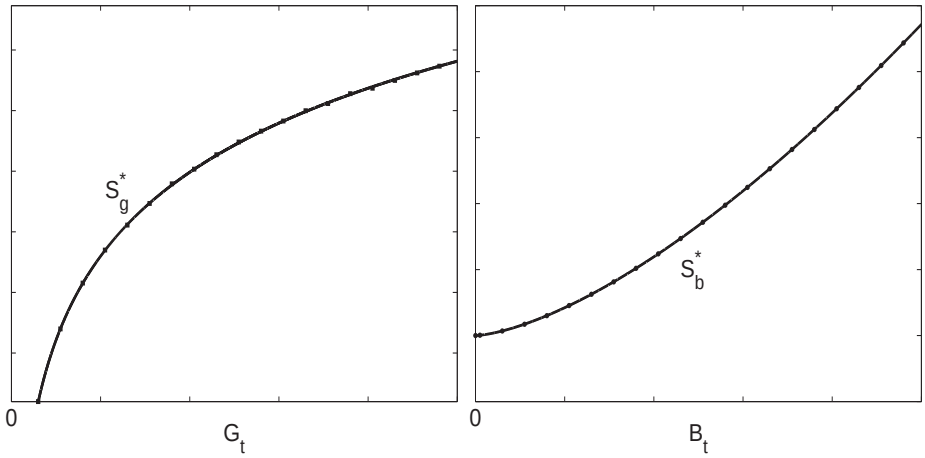


Figure 5.1: Calibrated scrap value functions: unbounded utility.

As an example, we graph the scrap value functions (5.7) in Figure 5.1 for a typical case. We see that S_g^* is concave and unbounded (left panel), and that S_b^* is convex, bounded from below, and unbounded from above (right panel). For S_g^* we need to estimate only ν_1; for S_b^* we need to estimate both ν_2 and q. Estimating these parameters as outlined in Sections 5.3 and 5.4, we obtain the two curves in Figure 5.1, where the circles represent the data generated from the infinite-horizon model.

5.6 Utility bounded from above but not from below

Consider the power utility function

$$U(x) = \frac{x^{1-\alpha} - 1}{1 - \alpha} \qquad (\alpha > 0). \tag{5.8}$$

In the previous section we considered the case $\alpha = 1$, that is, $U(x) = \log x$. Let us now consider the class $\alpha > 1$. Such utility functions are often used, in particular the case $\alpha = 2$ where $U(x) = 1 - 1/x$. Welfare is then bounded from above but not from below. The linear scrap value function does not share this feature. Hence we consider the nonlinear scrap value function (5.5), and choose the scrap value functions within the family of power functions. This gives

$$S_g^*(G_T) = -\zeta_1 G_T^{-p}, \qquad S_b^*(B_T) = \zeta_2 B_T^q$$

with $p > 0$, $q \geqslant 1$, $\zeta_1 > 0$, and $\zeta_2 > 0$. Then, S_g^* is increasing and strictly concave on $(0, \infty)$, and S_b^* is increasing and convex on $(0, \infty)$. Inserting these two functions in (5.5) we obtain

$$S_g^*(G_T) = -\frac{G_0}{p}\left(\frac{G_T}{G_0}\right)^{-p}, \qquad S_b^*(B_T) = \frac{B_0}{q}\left(\frac{B_T}{B_0}\right)^{q}, \tag{5.9}$$

where $p > 0$ and $q \geqslant 1$, and we have normalized ν_1 and ν_2 as before, so that they can be interpreted as scrap prices.

Let us also consider an alternative specification for S_b^*, namely

$$S_b^*(B_T) = \frac{\gamma_0 B_T^{q+r}}{1 + B_T^q} \qquad (q \geqslant 1,\ 0 < r < 1), \tag{5.10}$$

where

$$\gamma_0 = \frac{(1 + B_0^q)^2}{B_0^{q+r-1}(rB_0^q + q + r)}.$$

For B_T close to zero the function S_b^* behaves like B^{q+r} (convex) and for B_T close to ∞ it behaves like B^r (concave), and there is precisely one point where S_b^* turns from convex to concave. Hence, S_b^* is now increasing and convex-concave on $(0, \infty)$.

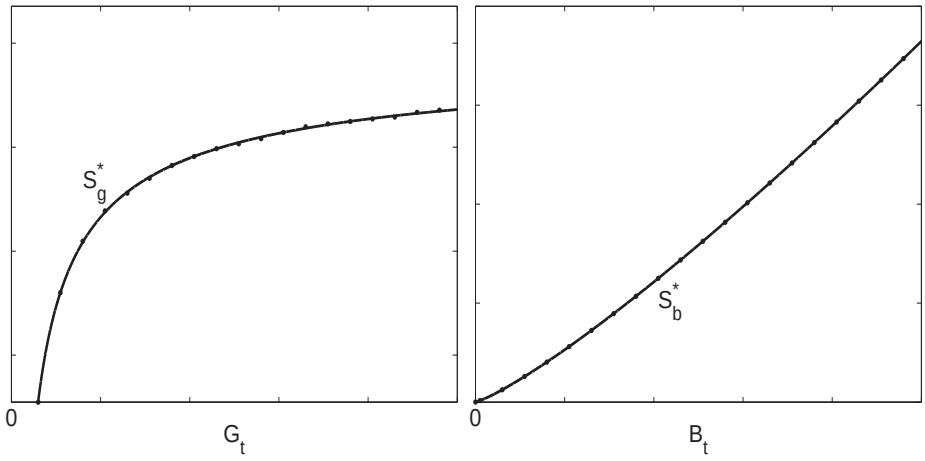


Figure 5.2: Calibrated scrap value functions: utility bounded from above.

The scrap value functions (5.9) are graphed in Figure 5.2. We see that S_g^* is concave, bounded from above, but unbounded from below (left panel), and that S_b^* is convex, bounded from below, but unbounded from above (right panel). We need to estimate the parameters (ν_1, p) for S_g^*, and (ν_2, q) for S_b^*. In this case it is clear from the data that the convex specification (5.9) for S_b^* fits the data better than the alternative convex-concave specification (5.10), but this depends of course on the case under investigation.

5.7 Utility bounded from below but not from above

This class of utility functions, which includes $U(x) = x^r$ for $0 < r < 1$ and $U(x) = \log(x + \lambda)$ for $\lambda > 0$, is not used as often as the other three classes. It is not difficult to find a suitable scrap value function for G_T, namely

$$S_g^*(G_T) = \frac{G_0^{1-r} G_T^r}{r} \qquad (0 < r < 1). \tag{5.11}$$

The function S_g^* is increasing and strictly concave on $(0, \infty)$.

It is more difficult to find a suitable scrap value function for B_T. As discussed in Section 5.2, when utility is bounded from below, as is the case in this and the next section, then S_b^* must be bounded from above, and this causes a problem because there are no increasing convex functions that are bounded from above. By Assumption 1, S_b^* is also bounded from below. Hence we are looking for a function S_b^* that is increasing, convex-concave, and bounded, in other words, we are looking for a general class of distribution functions. The Burr cumulative distribution function (Burr, 1942; Burr and Cislak, 1968), defined for $z > 0$ as

$$F(z) = 1 - (1 + (z/\lambda)^c)^{-p} \qquad (\lambda > 0,\, p > 0,\, c > 0),$$

is a three-parameter family of distribution functions with the property that many of the known distribution functions are special or limiting cases. It is therefore an appropriate function to approximate an unknown distribution function. The function F is increasing between $F(0) = 0$ and $F(\infty) = 1$, and

$$F''(z) < 0 \iff \left(\frac{z}{\lambda}\right)^c > \frac{c-1}{cp+1}.$$

Hence, F is concave when $0 < c \leqslant 1$ and convex-concave when $c > 1$. We therefore specify

$$S_b^*(B_T) = \gamma_1 \left[1 - (1 + (B_T/\lambda)^c)^{-p}\right] \qquad (\lambda > 0,\, p > 0,\, c > 1), \tag{5.12}$$

where

$$\gamma_1 = \frac{\lambda}{cp} \cdot \frac{(1 + (B_0/\lambda)^c)^{p+1}}{(B_0/\lambda)^{c-1}}.$$

The function S_b^* is then increasing and convex-concave on $(0, \infty)$.

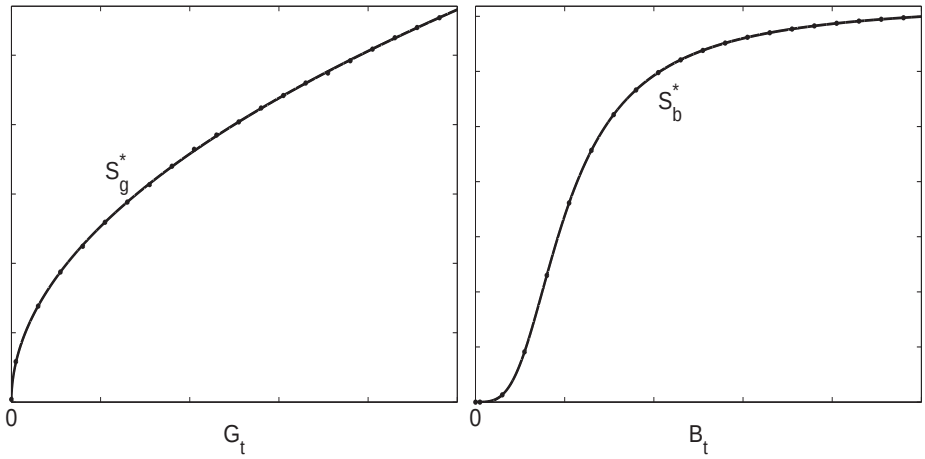


Figure 5.3: Calibrated scrap value functions: utility bounded from below.

The scrap value functions (5.11) and (5.12) are graphed in Figure 5.3. We see that S_g^* is concave, bounded from below, but not from above (left panel), and that S_b^* is convex-concave and bounded (right panel). For S_g^* we need to estimate (ν_1, r); for S_b^* we need to estimate (ν_2, λ, p, c).

5.8 Bounded utility

If we assume an exponential utility function

$$U(x) = 1 - e^{-\beta x} \qquad (\beta > 0) \tag{5.13}$$

or the 'Burr' utility function (Chapter 4)

$$U(x) = 1 - \left(\frac{\lambda}{x+\lambda}\right)^k \qquad (\lambda > 0,\, k > 0), \tag{5.14}$$

then welfare is bounded from above and from below. Hence we require S_g^* to be increasing, concave, and bounded; and S_b^* to be increasing, convex-concave, and bounded. Using again the flexibility of the Burr cumulative distribution function, we obtain

$$S_g^*(G_T) = \gamma_2 \left[1 - (1 + G_T/\lambda_1)^{-p}\right] \qquad (\lambda_1 > 0,\, p > 0) \tag{5.15}$$

and

$$S_b^*(B_T) = \gamma_3 \left[1 - (1 + (B_T/\lambda_2)^c)^{-q}\right] \qquad (\lambda_2 > 0,\, q > 0,\, c > 1), \tag{5.16}$$

where

$$\gamma_2 = \frac{\lambda_1}{p}\left(1 + G_0/\lambda_1\right)^{p+1}, \qquad \gamma_3 = \frac{\lambda_2}{cq} \cdot \frac{(1 + (B_0/\lambda_2)^c)^{q+1}}{(B_0/\lambda_2)^{c-1}}.$$

We see that S_g^* is increasing and concave on $(0, \infty)$, and that S_b^* is increasing and convex-concave on $(0, \infty)$.

The scrap value functions (5.15) and (5.16) are graphed in Figure 5.4. We see that S_g^* is concave and bounded (left panel), and that S_b^* is convex-concave and bounded (right panel). For S_g^* we need to estimate (ν_1, λ_1, p); for S_b^* we need to estimate (ν_2, λ_2, q, c).

5.9 Conclusion

In long-horizon dynamic stochastic models, such as discrete infinite-horizon welfare-maximizing problems, it is typically not possible to determine the optimal policy analytically, and numerical solutions may be computationally

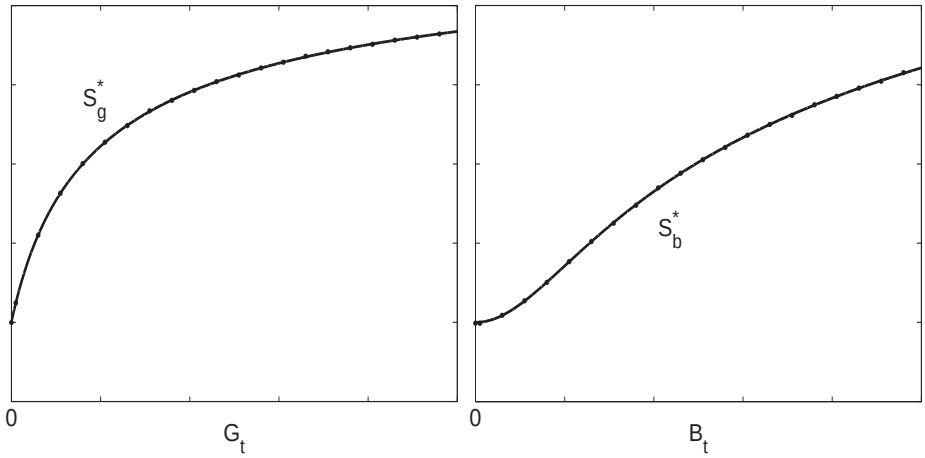


Figure 5.4: Calibrated scrap value functions: bounded utility.

expensive. In this chapter, we present a method which greatly reduces this computational effort.

Our method employs the idea of a scrap value function, which we estimate based on the deterministic version of the model. The form of the scrap value function depends on the form of the utility function, and this is explicitly taken into account by considering four types of utility function. A finer distinction does not appear to be useful. We also distinguish between state variables that are 'good' (like capital) and 'bad' (like pollution). In our analysis we have assumed a single random shock, but our method is easily generalized to multiple random shocks, as long as they appear before period T.

In our experience the estimated functions typically fit the data very well over the entire support, especially when the utility function belongs to the HARA class (as all our examples do), that is the class of utility functions with linear absolute risk tolerance $T(x) = -U'(x)/U''(x)$.

Chapter 6

Specification of variance matrices for panel data models

Abstract: Many regression models have two dimensions, say time ($t = 1, \ldots, T$) and households ($i = 1, \ldots, N$), as in panel data, error components, or spatial econometrics. In estimating such models we need to specify the structure of the error variance matrix Ω, which is of dimension $TN \times TN$. If TN is large, then direct computation of the determinant and inverse of Ω is not practical. In this chapter we define structures of Ω that allow the computation of its determinant and inverse, only using matrices of orders T and N, and at the same time allowing for heteroskedasticity, for household- or station-specific correlation, and for time-specific spatial correlation.

6.1 Introduction

We consider regression models with two dimensions, which we denote by T (say time) and N (say households or stations), such as

$$y_{it} = f_i(X_{it}, \beta_i) + u_{it} \qquad (i = 1, \ldots, N;\ t = 1, \ldots, T).$$

In estimating such models we need to specify the structure of the variance matrix Ω of the errors u_{it}, which will be of dimension $TN \times TN$. We shall assume that TN is so large that direct computation of its determinant and inverse is not practical. Thus we need to find a structure of Ω that allows the computation of its determinant and inverse, only using matrices of orders T and N but not TN. In this chapter we attempt to obtain maximum flexibility of the variance matrix under precisely this constraint. The flexibility that we aim for should allow for heteroskedasticity in the errors, for household- or station-specific correlation, and also for time-specific spatial correlation.

Problems of this nature arise in the panel data and error components literature; see Baltagi and Raj (1992), Baltagi (2001), and Arellano (2003) for useful reviews and historical details. They are also important in the closely related area of spatial econometrics; see Anselin (1988), Anselin and Bera (1998), Driscoll and Kraay (1998), Baltagi, Song, and Koh (2003), Baltagi, Song, Jung, and Koh (2007), and Kapoor, Kelejian, and Prucha (2007). The idea of introducing heteroskedasticity into error component models is discussed in Baltagi and Griffin (1988), and Li and Stengos (1994). Closest to our approach are the papers by Searle and Henderson (1979), and Wansbeek and Kapteyn (1982), who try to understand, like us, which class of variance matrices are appropriate for models with two dimensions.

We combine the errors in a matrix

$$U := \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1T} \\ u_{21} & u_{22} & \dots & u_{2T} \\ \vdots & \vdots & & \vdots \\ u_{N1} & u_{N2} & \dots & u_{NT} \end{pmatrix}, \tag{6.1}$$

and we define its T columns and N rows as

$$U = (u_1, u_2, \dots, u_T), \qquad U' = (\tilde{u}_1, \tilde{u}_2, \dots, \tilde{u}_N). \tag{6.2}$$

Letting $u := \text{vec}(U)$, we then have

$$\Omega := \text{var}(u) = \begin{pmatrix} \Omega_{11} & \Omega_{12} & \dots & \Omega_{1T} \\ \Omega_{21} & \Omega_{22} & \dots & \Omega_{2T} \\ \vdots & \vdots & & \vdots \\ \Omega_{T1} & \Omega_{T2} & \dots & \Omega_{TT} \end{pmatrix}, \tag{6.3}$$

where each of the submatrices is of order $N \times N$.

The simplest case is of course $\Omega = A \otimes B$, but this is usually not sufficiently general. The following result is often useful and is stated here separately, because of its importance and also because we will refer to it in the sequel. It is a special case of Lemma 2.2 of Magnus (1982).

LEMMA 6.1. Let A_k $(k = 1, 2, \dots, K)$ be symmetric idempotent matrices of order $T \times T$ and rank r_k satisfying $\sum_k A_k = I_T$, and let B_k $(k = 1, 2, \dots, K)$ be positive definite of order $N \times N$. Define $\Omega := \sum_k (A_k \otimes B_k)$ of order $TN \times TN$. Then, Ω is positive definite and its eigenvalues are the eigenvalues of $B_1, B_2, \dots, B_K$ with multiplicities $r_1, r_2, \dots, r_K$. Further,

$$|\Omega| = \prod_{k=1}^{K} |B_k|^{r_k}, \qquad \Omega^{-1} = \sum_{k=1}^{K} (A_k \otimes B_k^{-1}).$$

In Section 6.2 we make the simplifying assumption that the T variance matrices Ω_{tt} are free, but that the correlation matrices are constant. In Section 6.3 we consider the two error components model which is perhaps the main tool for panel data. We will see that considerably more flexibility is possible than previously utilized in the panel literature. In Section 6.4 we study the case where assumptions on the columns of Ω are combined with an independence assumption on (linear combinations of) the rows. In Section 6.5 we consider the three error components model and try and understand why this more general set-up does not lead to a more general specification. Section 6.6 concludes.

6.2 Constant correlation

Let us write the variance matrix Ω in terms of its correlation matrices

$$P_{st} := \Omega_{ss}^{-1/2} \Omega_{st} \Omega_{tt}^{-1/2},$$

so that Ω takes the form

$$\begin{pmatrix} \Omega_{11}^{1/2} & 0 & \dots & 0 \\ 0 & \Omega_{22}^{1/2} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \Omega_{TT}^{1/2} \end{pmatrix} \begin{pmatrix} I_N & P_{12} & \dots & P_{1T} \\ P_{21} & I_N & \dots & P_{21} \\ \vdots & \vdots & & \vdots \\ P_{T1} & P_{T2} & \dots & I_N \end{pmatrix} \begin{pmatrix} \Omega_{11}^{1/2} & 0 & \dots & 0 \\ 0 & \Omega_{22}^{1/2} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \Omega_{TT}^{1/2} \end{pmatrix}.$$

Suppose we like to keep maximum flexibility on the structure of the variance matrices Ω_{tt}, but that we are willing to assume that all correlation matrices are the same: $P_{st} = P$. In the special case where $P = 0$, this means that the error vectors $\{u_t\}$ are uncorrelated over time. The current assumption is more general. Also, with a obvious change of indices, we may assume that there is zero or constant correlation over households rather than over time.

The determinant and inverse of Ω can then be obtained from the following theorem.

THEOREM 6.1. Let the st-th block of Ω be defined as

$$\Omega_{st} := \begin{cases} B_t & \text{if } s = t, \\ B_s^{1/2} P B_t^{1/2} & \text{if } s \neq t, \end{cases}$$

where the B_t are all positive definite $N \times N$ matrices, and P is a symmetric $N \times N$ matrix whose eigenvalues are bounded by

$$\lambda_{\min}(P) > \frac{-1}{T-1}, \qquad \lambda_{\max}(P) < 1.$$

Define the two matrices

$$C_1 := I_N + (T-1)P, \qquad C_2 := I_N - P.$$

Then, Ω is positive definite, its determinant is given by

$$|\Omega| = |C_1||C_2|^{T-1} \prod_{t=1}^{T} |B_t|,$$

and the st-th block of Ω^{-1} by

$$\Omega^{st} := \begin{cases} \frac{1}{T} B_t^{-1/2} (C_1^{-1} + (T-1)C_2^{-1}) B_t^{-1/2} & \text{if } s = t, \\ \frac{1}{T} B_s^{-1/2} (C_1^{-1} - C_2^{-1}) B_t^{-1/2} & \text{if } s \neq t. \end{cases}$$

Proof: The matrix C_1 is positive definite because $\lambda_{\min}(P) > -1/(T-1)$, and C_2 is positive definite because $\lambda_{\max}(P) < 1$. Next, let

$$\bar{B} := \begin{pmatrix} B_1 & 0 & \dots & 0 \\ 0 & B_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & B_T \end{pmatrix}, \qquad \Omega_1 := \begin{pmatrix} I_N & P & \dots & P \\ P & I_N & \dots & P \\ \vdots & \vdots & & \vdots \\ P & P & \dots & I_N \end{pmatrix},$$

so that $\Omega = \bar{B}^{1/2} \Omega_1 \bar{B}^{1/2}$. Letting $p := \imath/\sqrt{T}$, where $\imath$ denotes the vector of ones, we can write Ω_1 more conveniently as

$$\Omega_1 = pp' \otimes C_1 + (I_T - pp') \otimes C_2.$$

Since pp' and $I_T - pp'$ are idempotent and sum to I_T, and since their ranks are 1 and $T-1$ respectively, we see from Lemma 6.1 that Ω_1 (and hence Ω)

is positive definite, and

$$|\Omega_1| = |C_1||C_2|^{T-1}, \qquad \Omega_1^{-1} = pp' \otimes C_1^{-1} + (I_T - pp') \otimes C_2^{-1}.$$

This implies that

$$|\Omega| = |\Omega_1||\bar{B}| = |C_1||C_2|^{T-1} \prod_{t=1}^{T} |B_t|,$$

and

$$\begin{aligned}\Omega^{-1} &= \bar{B}^{-1/2}\Omega_1^{-1}\bar{B}^{-1/2} = \bar{B}^{-1/2}(pp' \otimes C_1^{-1} + (I_T - pp') \otimes C_2^{-1})\bar{B}^{-1/2} \\ &= \bar{B}^{-1/2}\left(I_T \otimes C_2^{-1} + \frac{ii'}{T} \otimes (C_1^{-1} - C_2^{-1})\right)\bar{B}^{-1/2}.\end{aligned}$$

The result follows. □

6.3 Two error components

Although it is trivial to find the determinant and inverse of a simple Kronecker product $A \otimes B$, it is not so simple to do the same for the sum of two Kronecker products: $A_1 \otimes B_1 + A_2 \otimes B_2$. Only special cases allow explicit solutions, and the most general special case seems to be the following result.

THEOREM 6.2. Let A be a positive definite $T \times T$ matrix, a a nonzero $T \times 1$ vector, B_1 a positive semidefinite $N \times N$ matrix (possibly the null matrix), and B_2 a positive definite $N \times N$ matrix. Then the $TN \times TN$ matrix

$$\Omega := aa' \otimes B_1 + A \otimes B_2$$

is positive definite with

$$|\Omega| = |A|^N |B_2|^{T-1} |B_2 + (a'A^{-1}a)B_1|$$

and

$$\Omega^{-1} = \frac{A^{-1}aa'A^{-1}}{a'A^{-1}a} \otimes \big(B_2 + (a'A^{-1}a)B_1\big)^{-1} + \left(A^{-1} - \frac{A^{-1}aa'A^{-1}}{a'A^{-1}a}\right) \otimes B_2^{-1}.$$

Proof: Let

$$\alpha^2 := a'A^{-1}a, \qquad p := A^{-1/2}a/\alpha, \qquad C := B_2 + \alpha^2 B_1.$$

Then,

$$\Omega = aa' \otimes B_1 + A \otimes B_2 = (A^{1/2} \otimes I_N)\Omega_1(A^{1/2} \otimes I_N),$$

where

$$\begin{aligned} \Omega_1 &:= A^{-1/2}aa'A^{-1/2} \otimes B_1 + I_T \otimes B_2 \\ &= pp' \otimes \alpha^2 B_1 + I_T \otimes B_2 \\ &= pp' \otimes C + (I_T - pp') \otimes B_2. \end{aligned}$$

Since pp' and $I_T - pp'$ are idempotent matrices which sum to I_T, and B_2 and C are both positive definite, it follows from Lemma 6.1 that Ω_1 is positive definite, and that

$$|\Omega_1| = |B_2|^{T-1}|C|, \qquad \Omega_1^{-1} = pp' \otimes C^{-1} + (I_T - pp') \otimes B_2^{-1}.$$

The determinant and inverse of Ω now follow easily. □

Special cases of Theorem 6.2 have been studied in the literature. Thus, Baltagi, Song, and Koh (2003) consider the case where $a = \imath$, $B_1 = \sigma_1^2 I_N$,

$A = \sigma_2^2 I_T$, and $B_2 = ((I_N - \lambda W)'(I_N - \lambda W))^{-1}$, which arises from a structure with random regional effects and spatially correlated errors. In particular, it is assumed that

$$u_{it} = v_i + \epsilon_{it}, \qquad \epsilon_t = \lambda W \epsilon_t + e_t,$$

where v_i and ϵ_{it} are independent, $\text{var}(v_i) = \sigma_1^2 I_T$, and $\text{var}(e_t) = \sigma_2^2 I_N$. In a sequel paper, Baltagi, Song, Jung, and Koh (2007) assume in addition that the remainder term e_t follows a first-order autoregressive process by defining A to be the familiar AR(1) variance matrix. Theorem 6.2 shows that considerably more generality is still feasible.

6.4 Weak row-independence

Some form of independence must be assumed in order to get a manageable variance matrix Ω. One often wants to make assumptions on the columns of Ω while it is reasonable to assume independence of (some transformation of) the rows. The following result is somewhat related to Kapoor, Kelejian, and Prucha (2007), who also wish to combine household- or station-specific autocorrelation with time-specific spatial correlation.

THEOREM 6.3. Let the error vectors $u_1, u_2, \ldots, u_T$ be generated by $u_t = B_t \epsilon_t$, where the vectors $\tilde{\epsilon}_1, \tilde{\epsilon}_2, \ldots, \tilde{\epsilon}_N$ are independently distributed with mean zero and variance $\text{var}(\tilde{\epsilon}_i) = A_i$. Then,

$$|\Omega| = \prod_{t=1}^{T} |B_t' B_t| \prod_{i=1}^{N} |A_i|,$$

and, if all A_i and B_t are nonsingular, Ω is positive definite, and the st-th block of Ω^{-1} is given by

$$\Omega^{st} = (B_s')^{-1} \operatorname{diag}\left(A_1^{st}, A_2^{st}, \ldots, A_N^{st}\right) B_t^{-1}.$$

where A_i^{st} denotes the st-th element of A_i^{-1}.

Proof: Let the matrix $N \times T$ matrix U be defined as in (6.1) with T columns $u_1, u_2, \ldots, u_T$ and N rows $\tilde{u}_1', \tilde{u}_2', \ldots, \tilde{u}_N'$, and let E be defined similarly. Further, let

$$\bar{A} := \begin{pmatrix} A_1 & 0 & \ldots & 0 \\ 0 & A_2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & A_N \end{pmatrix}, \qquad \bar{B} := \begin{pmatrix} B_1 & 0 & \ldots & 0 \\ 0 & B_2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & B_T \end{pmatrix}.$$

Finally, let K denote the $NT \times NT$ commutation matrix which transforms $\mathrm{vec}(E)$ into $\mathrm{vec}(E')$, so that $K\,\mathrm{vec}(E) = \mathrm{vec}(E')$; see Magnus and Neudecker (1988, Section 3.7) and Magnus (1988) for further details. Note that K is a permutation matrix, hence orthogonal: $K' = K^{-1}$. Then,

$$u := \mathrm{vec}(U) = \bar{B}\,\mathrm{vec}(E) = \bar{B}K'\,\mathrm{vec}(E'),$$

so that

$$\Omega := \mathrm{var}(u) = \bar{B}K'\,\mathrm{var}(\mathrm{vec}\,E')K\bar{B}' = \bar{B}K'\bar{A}K\bar{B}'.$$

This implies that

$$|\Omega| = \prod_{t=1}^{T} |B_t'B_t| \prod_{i=1}^{N} |A_i|,$$

and

$$\Omega^{-1} = (\bar{B}')^{-1}K'\bar{A}^{-1}K\bar{B}^{-1}.$$

In order to obtain explicit expressions for the blocks of Ω^{-1}, we let p_i be the i-th column of I_N and q_t the t-th column of I_T. Then we can write

$K = \sum_i \sum_t (p_i q_t' \otimes q_t p_i')$, see Magnus (1988, Theorem 3.2). Hence,

$$\begin{aligned}
K'\bar{A}^{-1}K &= \sum_{i,j,h=1}^{N} \sum_{s,t=1}^{T} (q_s p_i' \otimes p_i q_s')(p_h p_h' \otimes A_h^{-1})(p_j q_t' \otimes q_t p_j') \\
&= \sum_{i,j,h=1}^{N} \sum_{s,t=1}^{T} (q_s p_i' p_h p_h' p_j q_t') \otimes (p_i q_s' A_h^{-1} q_t p_j') \\
&= \sum_{i=1}^{N} \sum_{s,t=1}^{T} (q_s q_t') \otimes (p_i q_s' A_i^{-1} q_t p_i') \\
&= \sum_{s,t=1}^{T} (q_s q_t') \otimes \sum_{i=1}^{N} A_i^{st} p_i p_i' \\
&= \sum_{s,t=1}^{T} (q_s q_t') \otimes \operatorname{diag}\left(A_1^{st}, A_2^{st}, \ldots, A_N^{st}\right),
\end{aligned}$$

from which the blocks Ω^{st} follow directly. □.

We notice that in Theorem 6.3 all three objectives have been realized. There is heteroskedasticity (through A_i), there is spatial correlation (through B_t, for example by specifying $\epsilon_t = W\epsilon_t + u_t$, so that $B_t = I_N - W$, in this case constant), and there is household-specific correlation (also through A_i).

6.5 Three error components

One may wonder whether a useful extension from two error components to three error components is possible. It turns out that this is not the case, and we briefly investigate why this is so.

A three error component model would consist of three independent errors leading to a sum of three variance matrices. The most general case seems to be

$$\Omega := aa' \otimes B + A \otimes bb' + \tau^2 A \otimes B,$$

where A is a positive definite $T \times T$ matrix, a a nonzero $T \times 1$ vector, B a positive definite $N \times N$ matrix, and b a nonzero $N \times 1$ vector. The $TN \times TN$

matrix Ω is positive definite, and we can calculate its determinant and inverse by writing

$$\Omega = (A^{1/2} \otimes B^{1/2})\Omega_1(A^{1/2} \otimes B^{1/2}),$$

where

$$\begin{aligned}\Omega_1 &:= A^{-1/2}aa'A^{-1/2} \otimes I_N + I_T \otimes B^{-1/2}bb'B^{-1/2} + \tau^2 I_T \otimes I_N \\ &= \alpha^2 pp' \otimes I_N + \beta^2 I_T \otimes qq' + \tau^2 I_T \otimes I_N = \sum_{h=1}^{4} \omega_h^2 J_h\end{aligned}$$

with

$$\alpha^2 := a'A^{-1}a, \qquad p := A^{-1/2}a/\alpha, \qquad \beta^2 := b'B^{-1}b, \qquad q := B^{-1/2}b/\beta$$

and

$$\begin{aligned}\omega_1^2 &:= \alpha^2 + \beta^2 + \tau^2, & J_1 &:= pp' \otimes qq', \\ \omega_2^2 &:= \alpha^2 + \tau^2, & J_2 &:= pp' \otimes (I_N - qq'), \\ \omega_3^2 &:= \beta^2 + \tau^2, & J_3 &:= (I_T - pp') \otimes qq', \\ \omega_4^2 &:= \tau^2, & J_4 &:= (I_T - pp') \otimes (I_N - qq').\end{aligned}$$

Since pp' and qq' are idempotent matrices of rank one, the matrices J_h are all idempotent with rank $\text{rank}(J_h) = r_h$. Since also $\sum_h J_h = I_{TN}$, Lemma 6.1 applies again.

It may seem that we have obtained a generalization of the two error components structure, but in fact we have not. This is easily seen by writing Ω in the form

$$\Omega := aa' \otimes B + A \otimes (bb' + \tau^2 B).$$

In fact, this form of the three error component structure is less rather than more general than the two error component structure considered in Theorem 6.2.

6.6 Conclusions

In this chapter we have presented four possible ways in which the $TN \times TN$ variance matrix of panel data models can be specified, allowing for maximum flexibility under the constraint that the determinant and inverse of the variance matrix can be calculated from matrices of orders $T \times T$ and $N \times N$ only. We conclude that much more generality is possible than is typically applied in panel data specifications.

Chapter 7

Efficient GMM estimation with a general missing data pattern

Abstract: This chapter considers GMM estimation from a random sample of incomplete observations. For each observation, certain components of the moment function may be unavailable. We propose an estimator for an arbitrary set of regular moment conditions and a general missing data pattern. The estimator is consistent and asymptotically efficient under an assumption that is weaker than missing completely at random. It can be interpreted as the optimal linear combination of subsample GMM estimators. Because of this linearity, the computational burden and the small-sample performance of the estimator are comparable to the full-data estimator. We also propose an inverse probability weighted version of the estimator that is consistent when selection is on observables. Applications to multivariate mean estimation, instrumental variable estimation, and dynamic panel data estimation demonstrate the efficiency gain with respect to existing missing data methods. We also discuss how the results can be used to optimize data collection for measuring consumer confidence.

7.1 Introduction

Missing data affect the majority of empirical studies in economics. In a survey of empirical research in top economics journals, Abrevaya and Donald (2010) find that missing data occurs in 40% of the publications. In 70% of these cases, a complete-case estimator is used. A complete-case estimator discards all incomplete observations. This is inefficient if the incomplete observation contain information about the parameter of interest.

The main contribution of this chapter is to introduce an estimation procedure that efficiently combines information from complete and incomplete observations. Interest is in GMM estimation of a finite-dimensional parameter with a random sample. Our procedure can be applied to two-step, iterative and continuous updating GMM estimators. We do not impose restrictions on the missing data *pattern*, which means that the data can be incomplete in an arbitrary way. In terms of the missing data *mechanism*, we assume that there is no selection or that selection is on observables.

In many econometric models, observations that are incomplete can still be informative. To see this, consider instrumental variables estimation and dynamic panel data models. First, a linear instrumental variable model with one endogenous variable X and two instruments $Z = (Z_1, Z_2)$ is given by the equation $y = X\beta_0 + \epsilon$ and the conditional mean assumption $\mathrm{E}(\epsilon|\, Z) = 0$. Estimation of the parameter β_0 is based on the unconditional moment condition $\mathrm{E}(Z\epsilon) = 0$. Assume that for each observation both y and X are observed. An observation with no measurement for instrument Z_2 will still be useful if the other instrument, Z_1, is observed. To see this, consider the subsample of all observations for which only (y, X, Z_1) is observed. This subsample is informative, since we can use it to estimate β_0 using the moment condition $\mathrm{E}(Z_1\epsilon) = 0$. Missing instruments are common in empirical research, see for example Levitt (2002), who uses the number of firefighters and the number of city workers as instruments to estimate the effect of police on crime. Not all cities provide information about the number of firefighters in each year, and data on the number of city workers is available for yet another subsample. Another example can be found in Rodrik et al. (2004), who investigate the effect

	Missing components			
	None	$y_{i,1}$	$y_{i,4}$	$(y_{i,1}, y_{i,4})$
$y_{i,1}\Delta\epsilon_{i,3}$	X	·	X	·
$y_{i,1}\Delta\epsilon_{i,4}$	X	·	·	·
$y_{i,1}\Delta\epsilon_{i,5}$	X	·	·	·
$y_{i,2}\Delta\epsilon_{i,4}$	X	X	·	·
$y_{i,2}\Delta\epsilon_{i,5}$	X	X	·	·
$y_{i,3}\Delta\epsilon_{i,5}$	X	X	·	·

Table 7.1: Missing data patterns for dynamic panel data estimation using the estimator in Arellano and Bond (1991), $T = 5$.

of institutions and geography on economic growth by using trade predictions and settler mortality rates as instruments that are sometimes unobserved.

Dynamic panel data models provide a second example of incomplete, informative observations. Interest is in the autoregressive parameter ρ in

$$y_{i,t} = \alpha_i + \rho y_{i,t-1} + \epsilon_{i,t},\ 2 \leqslant t \leqslant T.$$

Arellano and Bond (1991) propose an estimator that is based on the absence of serial correlation in the error terms, which implies the moment conditions

$$\mathrm{E}(y_{i,t-s}\Delta\epsilon_{i,t}) = 0,\ t \geqslant 3,\ s \geqslant 2.$$

Table 7.1 illustrates the relationship between the incompleteness of an observation and the extend to which that observation contributes to the sample moment. In Table 7.1, we consider the case of $T = 5$ time periods and six moment conditions. If $y_{i,1}$ is missing, observation i still contributes to three sample moments. If $y_{i,4}$ is missing, only one component of the moment function can be evaluated. More generally, the estimator proposed in this chapter can efficiently accommodate static and dynamic panel data models with unbalanced panels with different starting points, endpoints, and any combination of gaps.

Standard approaches that, in contrast to the complete-case estimator, use

all available information can still be inefficient. One such approach is the available-case estimator, which replaces missing moments by zeros before applying the full data estimation procedure. The available-case estimator is consistent if there is no selection. In the instrument example, available-case estimation corresponds to replacing the missing instruments by zero. For the dynamic panel data example, it corresponds to the procedure suggested in Arellano and Bond (1991, p. 281).

The key to efficient estimation is to split the random sample in subsamples based on the missing data pattern. If two instruments are available, we can distinguish three subsamples: observations with measurements on both instruments are placed in the first subsample; observations with only the first instrument available are placed in the second subsample; the third subsample contains the observations that only have measurements on the second instrument. In the absence of selection, β_0 can be estimated using each subsample. Using efficient GMM in each subsample yields three consistent estimators of β_0. Any weighted average of these estimators is again a consistent estimator of β_0. The complete-case estimator assigns full weight to the estimator from the first subsample. The available-case estimator assigns equal weight to each estimator. We show that there exist optimal weights that minimize the asymptotic variance of the estimator.

The procedure is shown to be consistent under an assumption that is weaker than missing completely at random (MCAR). MCAR requires that the data are fully independent of the missing data indicator, and we only require that the moment condition holds conditional on the missing data indicator. Under this assumption, the estimator is asymptotically efficient in the sense that it attains the semiparametric efficiency bound. Furthermore, the computational and small sample properties are close to those of the full data estimator, since the minimization problem for the missing data estimator is linear in full data problems.

After introducing notation in Section 7.2, we show in Section 7.3 that the procedure using subsamples can be generalized to parameter vectors and that the parameter does not need to be identifiable in each subsample. Section 7.4 considers the special case where the parameter is identified in each subsample,

as discussed above. In Section 7.5, we show that our estimation procedure can be extended to a generalized inverse probability weighting estimator in order to deal with selection on observables. Again, we will work under an assumption that is weaker than the typical missing at random (MAR) assumption. Section 7.6 gives some examples, and show that substantial efficiency gains over standard approaches are possible. Section 7.7 concludes. Proofs can be found in the Appendix 7.A.

This chapter is not concerned with univariate regression methods. As soon as an observation is incomplete it will contribute to none of the sample moments and is therefore uninformative in our framework. The same holds for univariate instrumental variables case with missing dependent or endogenous variables. More generally, we are not concerned with situations in which each observation contributes either to all, or to none of the sample moments. For this case there is a vast literature that addresses efficient and robust estimation under MAR. This literature was initiated by Robins et al. (1994) and is still active, with recent contributions by Wang et al. (2004), Wooldridge (2007), Chen et al. (2008), Graham (2010) and Graham et al. (2010). Extending this literature to a general missing data pattern is theoretically and computationally challenging, see for example Tsiatis (2006, p. 255).

Finally, some papers consider specific GMM settings or specific missing data patterns. The static panel data setting is investigated by Chen et al. (2010). Abowd et al. (2001) allow for attrition in a dynamic panel data model. Instrumental variables estimation with missing instruments is discussed in Abrevaya and Donald (2010) and Mogstad and Wiswall (2010). Verbeek and Nijman (1992) study a static panel data setting and exploit the existence of different missing data patterns to test for selectivity bias.

7.2 Sample moments for missing data

We introduce notation for general missing data patterns, and discuss how a missing data pattern for X implies which subset of the components of a moment function $h(X, \theta)$ can be evaluated. We introduce an assumption about

the missing data mechanism, MI, which is a mean independence version of missing completely at random. MI is a sufficient condition for the complete- and available-case methods. In Section 7.3, we consider estimation under MI for data with a general missing data pattern.

7.2.1 Missing data patterns in GMM estimation

There are 2^d ways in which the components of a random vector $X \in \mathbb{R}^d$ can be missing, since each component is either missing or not. For a given model, the number of possible patterns is J_x, which can be smaller than 2^d when some patterns are ruled out by design. We use a diagonal selection matrix $S^x \in \mathbb{R}^{d\times d}$ to describe a missing data pattern. Such a matrix has kth diagonal entry equal to 1 if and only if the kth component of X is observed, that is:

$$(S^x)_{k_1,k_2} = \begin{cases} 1 & \text{if } k_1 = k_2 \text{ and component } k_1 \text{ is observed for pattern } j, \\ 0 & \text{otherwise.} \end{cases}$$

The J_x diagonal selection matrices $S_j^x,\ j = 1,\dots,J_x$, describe the missing data patterns. The missing data indicator $R^x \in \mathbb{R}^{d\times d}$ is a random matrix that captures which components of X are missing and takes values $S_j^x,\ 1 \leqslant j \leqslant J_x$.

In GMM estimation, a parameter of interest $\theta_0 \in \Theta \subset \mathbb{R}^p$ is defined through the moment conditions $\mathrm{E}(h(X,\theta_0)) = 0$, with moment function $h : \mathbb{R}^d \times \Theta \to \mathbb{R}^q$. If an observation is incomplete, only a subset of the components of the moment function is observable. A missing data pattern represented by S^x implies a missing moment pattern, which we describe by a diagonal selection matrix $S \in \mathbb{R}^{q\times q}$. As such, S describes a missing moment pattern for h in the same way that S^x describes a missing data pattern for X. The number of missing data patterns is greater than or equal to the number of missing moment patterns J, because different values for R^x can imply the same value for R. The missing moment indicator R takes values $S_j,\ 1 \leqslant j \leqslant J$. Let $p_j = \mathrm{P}(R = S_j)$ be the probability that missing moment pattern j occurs.

ASSUMPTION 7.1 (FULL-RANK). The probability of observing pattern j is positive, $p_j > 0$, for each $1 \leqslant j \leqslant J$ and $\operatorname{rank}\left(\sum_{j=1}^{J} S_j\right) = q$.

The restriction of positive probability is not restrictive, since we can eliminate patterns that occur with zero probability. The second restriction ensures that each component of the moment function is observed with positive probability.

7.2.2 Missing completely at random

Typically, three assumptions about the missing data mechanism are distinguished: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). For a detailed discussion of these concepts, see Little and Rubin (2002, Chapter 1). MCAR is the most restrictive assumption. Let $\perp$ denote statistical independence.

ASSUMPTION 7.2 (MCAR). $X \perp R^x$.

ASSUMPTION 7.3 (IID1). $(R_i^x, R_i^x X_i,\ 1 \leqslant i \leqslant n)$ is a random sample of size n from (R, RX).

Assumption MCAR requires that whether or not a random variable is observed is independent of the realization. For a moment function h, MCAR implies that $h(X, \theta) \perp R$ for each $\theta \in \Theta$ because $h(X, \theta)$ depends on X and not on R^x, while R is determined by R^x. This implies the following MCAR-like mean independence condition:

ASSUMPTION 7.4 (MI). $\mathrm{E}\left(Rh(X, \theta_0) \mid R = S_j\right) = 0$ for each $1 \leqslant j \leqslant J$.

This assumption requires the observable moment conditions to hold regardless of the missing data pattern. To demonstrate the difference between MCAR and MI, consider the univariate linear regression model, $y_i = \beta x_i + \epsilon_i,\ \mathrm{E}(\epsilon_i | X_i) = 0$. In Figure 1, we present the regression line and some simulated data. A cross represents an observation that is missing, $r_i = 0$, and a dot represents an observation that is complete, $r_i = 1$. The sample can be split in two groups: those with low x_i and those with high x_i. In terms

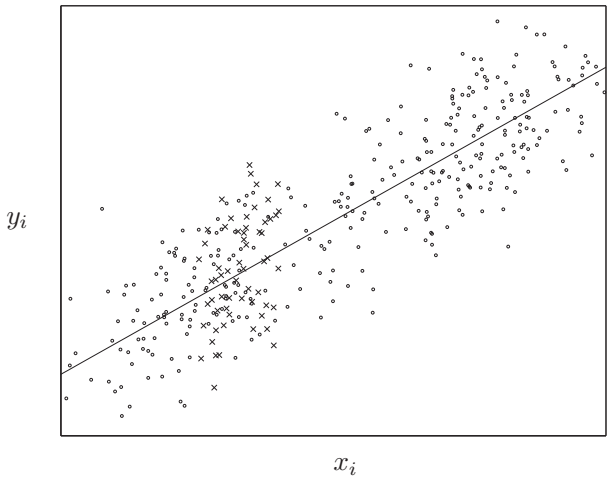


Figure 7.1: MI, not MCAR. Simulated data for a univariate regression model. A cross represents a missing observation; a dot represents a complete observation.

of deviation from the regression line, the data are arbitrarily missing in the sense that the estimator that uses the missing data has the same expectation as the estimator that uses the complete data. However, the situation in Figure 7.1 does not satisfy MCAR: an observation in the low group has a positive probability of being missing, while an observation in the high group is always complete, so $\mathrm{P}(r = 1 \mid X \text{ low}) \neq \mathrm{P}(r = 1 \mid X \text{ high})$. The data are MI, since $\mathrm{E}(x_i\epsilon_i | r = 1) = \mathrm{E}(x_i\epsilon_i | r = 0) = 0$. If we strengthened MI to include independence of the variance, or mean independence at values of the parameter other than the true value of β, MI would not be satisfied in this example: $\mathrm{var}(x_i\epsilon_i | r_i = 1) > \mathrm{var}(x_i\epsilon_i | r_i = 0)$, and $\mathrm{E}(x_i(y_i - (\beta + 1)x_i | r_i)) = \mathrm{E}(x_i\epsilon_i | r_i) - \mathrm{E}(x_i^2 | r_i) = -\mathrm{E}(x_i^2 | r_i) \neq \mathrm{E}(x_i^2)$.

The complete-case approach and the available-case approach are two popular ways to deal with missing data. Both methods are consistent under MI. The complete-case estimator is common in empirical work and is the default approach for most statistical packages. A complete-case estimator uses only complete observations. Let $S_1 = I_q$, so that all components of h can be evaluated for observations with missing data pattern 1. Then, the complete-case

sample moment for $\mathrm{E}(h(X,\theta))$ is

$$h_{\mathrm{cc},n}(\theta) = \frac{1}{n_1} \sum_{i \in G_1} R_i h(X_i, \theta),$$

where G_j is the subsample for which $R_i = S_j$ and n_j is the number of observations in subsample G_j, $1 \leqslant j \leqslant J$. A complete-case GMM estimator is based on the complete-case sample analog.

The available-case approach uses all the available data. For each component of the moment function it uses all the observations for which that component is observed. The available-case sample moment is

$$h_{\mathrm{ac},n}(\theta) = \frac{1}{n} \hat{\bar{R}}^{-1} \sum_{i=1}^{n} R_i h(X_i, \theta),$$

where the inverse of $\hat{\bar{R}} = \sum_{j=1}^{J} (n_j/n) S_j$ is used to divide each component of the sum by the number of observations that actually contribute.

In Section 7.3 we consider GMM estimation under MI, and we find an estimator that is asymptotically efficient under MI. In Section 7.5, we consider GMM estimation under a mean independence version of MAR.

7.3 GMM estimation

We are interested in estimating a parameter θ_0 that is defined through the moment conditions $\mathrm{E}(h(X,\theta_0)) = 0$. Given a complete data set, we would use the optimal GMM estimator. We construct a class of estimators that are consistent under MI. We show that the asymptotic variance of an optimal estimator in this class achieves the semiparametric efficiency bound for θ_0 under MI. The results in this section are a natural generalization of the properties of the optimal full-data GMM estimator to the optimal GMM estimator with a general missing data pattern. In Section 7.4, we consider a special case where the parameter can be estimated using the observations for an arbitrary pattern only. In Section 7.5, we allow the missing data indicator to depend

on observable random variables. We provide examples of the estimator in this section in Section 7.6. All the proofs are in Appendix 7.A.

7.3.1 GMM with missing data

We are interested in a parameter $\theta_0 \in \Theta \subset \mathbb{R}^p$ that is defined through a moment function $h : \mathbb{R}^d \times \Theta \to \mathbb{R}^q$ for which the following is assumed:

ASSUMPTION 7.5 (MI). $\mathrm{E}\,(Rh(X, \theta_0) \mid R = S_j) = 0$ for each $1 \leqslant j \leqslant J$.

ASSUMPTION 7.6 (IDENTIFICATION). For any $\theta \in \Theta$ for which $\theta \neq \theta_0$, there exists at least one pattern j for which $\mathrm{E}\,(h(X, \theta) | \, R = S_j) \neq 0$.

ASSUMPTION 7.7 (IID1). $(R_i^x, R_i^x X_i,\ 1 \leqslant i \leqslant n)$ is a random sample of size n from (R, RX).

A GMM estimator for θ_0 for complete data is defined as the minimizer over Θ of

$$\left(\sum_{i=1}^{n} h(X_i, \theta)\right)' W(n) \left(\sum_{i=1}^{n} h(X_i, \theta)\right), \tag{7.1}$$

for some arbitrary symmetric positive definite matrix $W(n)$. Since $h(X_i, \theta)$ is not observed for each i, this estimator is not feasible. For completeness, we restate the assumption about the available data and the missing data mechanism.

Let $h_{n,j}(\theta)$ be the sample moment for subsample $G_j = \{i : R_i = S_j\}$,

$$h_{n,j}(\theta) = (1/n_j) \sum_{i \in G_j} R_i h(X_i, \theta).$$

We define a GMM estimator for missing data as the minimizer of the modification of the full-data objective function (7.1),

$$\hat{\theta}_{W(n)} = \mathrm{argmin}_{\theta \in \Theta} \sum_{j=1}^{J} h_{n,j}(\theta)' W_j(n) h_{n,j}(\theta). \tag{7.2}$$

A GMM estimator for missing data minimizes the sum of weighted subsample moments instead of weighted sample moments. Complete-case and available-case estimators can be obtained as special cases. If pattern 1 is the complete-data pattern, $S_1 = I_q$, a complete-case estimator is obtained by setting $W_1(n) = W_{\text{cc},n}$ and $W_j(n) = 0_q,\ j > 1$, where $W_{\text{cc},n}$ can be chosen optimally. The available-case estimator follows from setting $W_j(n) = S_j W_{\text{ac}}(n) S_j$ for each $j = 1, \ldots, J$, where $W_{\text{ac}}(n)$ can be chosen optimally. By construction, our estimator will be at least as efficient as the complete-case and available-case estimators. The examples in Section 7.6 demonstrate that the efficiency gain is substantial.

The asymptotic distribution of the estimator $\hat{\theta}_{W(n)}$ requires the assumptions stated below.

Assumption 7.8 (finite-Ω_j). For each j, $\operatorname{var}(h(X, \theta_0) \mid R = S_j) = \Omega_j < \infty$, where $1 \leqslant j \leqslant J$.

The finite-Ω_j assumption is not compatible with mcar because finite-Ω_j allows the conditional variance of the moment function to depend on the missing data pattern.

Assumption 7.9 (derivative). (i) For each x, the moment function $h(x, \cdot)$ is continuously differentiable on Θ; (ii) for each pattern j let the $q \times p$ matrix $D_j(\theta) = \mathrm{E}(\partial h(X, \theta_0)/\partial\theta' \mid R)$ be uniformly bounded, in the sense that $\sup_{\theta \in \Theta} \|D_j(\theta)\| < \infty$, where $\|D_j\| = \operatorname{tr}(D_j' D_j)^{1/2}$; (iii) for each pattern j, $\operatorname{rank}(D_j) = p$.

Assumption 7.10 (regularity). (i) The parameter space Θ is compact and θ_0 is in the interior of Θ; (ii) the moment function is bounded in absolute mean:

$$\sup_{\theta \in \Theta} \mathrm{E}(|h(X, \theta)|) < \infty;$$

(iii) for each subsample, the sequence of GMM weights $(W_j(n),\ n \in \mathbb{N})$ satisfies $S_j W_j(n) S_j = W_j(n)$ and converges to a positive semidefinite matrix, W_j,

with $\text{rank}(W_j) = \text{rank}(S_j)$; (iv) the distribution of X conditional on $R = S_j$, represented by density $f_j(x)$, does not depend on θ

All conditions are standard GMM assumptions, except for REGULARITY(iii), which sets the submatrix of W_j that corresponds to $S_j = 0$ equal to zero, and requires the remaining submatrix to be positive definite, and (iv) which is satisfied the parameters of the process generating the data X to be seperate from those that generate the missings R. Compare the assumption of ignorability in Little and Rubin (2002).

THEOREM 7.1. Under assumptions MI, IDENTIFICATION, IID1, FULL-RANK, FINITE-Ω_j, DERIVATIVE, and REGULARITY, we have that, as $n \to \infty$,

$$\sqrt{n}\left(\hat{\theta}_{W(n)} - \theta_0\right) \xrightarrow{d} N\left(0, B^{-1}\left(\sum_{j=1}^{J} \frac{1}{p_j} D_j' W_j (S_j \Omega_j S_j) W_j D_j\right) B^{-1}\right),$$

where

$$B = \sum_{j=1}^{J} p_j D_j' (S_j \Omega_j S_j)^+ D_j. \tag{7.3}$$

Proof. The proof is given in Appendix 7.A. It involves converting the conditional moment restrictions in MI to an augmented set of Jq unconditional moment conditions. The expression in (7.2) can be seen as a weighted sample analog to this set of unconditional moment conditions. The double sum appears because we have a random sample, which implies independent subsamples. Then, we show that this is a standard GMM situation. □

The asymptotic variance can be minimized by setting each W_j equal to $W_j^* = p_j (S_j \Omega_j S_j)^+$. Note that this reduces to the familiar optimal weighting matrix if $J = 1$, $p_1 = 1$, and $S_1 = I_q$. The estimator that uses weighting matrices $W^*(n) = (W_1^*(n), \ldots, W_J^*(n))$ is denoted $\hat{\theta}_n^*$ and has limiting distribution

$$\sqrt{n}(\hat{\theta}_{W^*(n)} - \theta_0) \xrightarrow{d} N(0, B^{-1}). \tag{7.4}$$

This is an extension of the familiar result on optimal GMM: the weighting matrix for each subsample moment is proportional to the inverse of the relevant part of the variance matrix.

REMARK 7.1. The conditional moment assumptions MI can be viewed as a restriction on the conditional densities $f_j(x)$. Starting from a situation with no missings and $\mathrm{E}(h(X,\theta_0)) = 0$, we only allow conditional densities $f_j(x)$ that imply that the conditional expectation in the subpopulation is 0 when the parameter is equal to the true value that applies to the population. Here, we do not make explicit which conditional densities allow for. In general the set of compatible conditional densities will depend on h.

REMARK 7.2. It is possible to formulate identification conditions that are sufficient but not necessary for IDENTIFICATION. A useful condition is that identification in one subsample implies IDENTIFICATION. If there exists a subsample j such that $\mathrm{E}\left(Rh(X,\theta)|R = S_j\right) = 0 \Leftrightarrow \theta = \theta_0)$ then IDENTIFICATIONis satisfied.

REMARK 7.3. Replacing the variance matrices Ω_j and the derivative matrices D_j by consistent estimators leaves the asymptotic distribution of $\hat{\theta}_n^*$ unchanged.

REMARK 7.4. The GMM estimator based on the modified objective function is computationally slightly more expensive than the full-data sample moment. The only additional computational burden comes from determining J, rather than 1, optimal matrix weights, for which an analytical expression is available, and sorting the n observations into J groups.

7.3.2 Semiparametric efficiency bound

The model defined by MI and IID1 is a semiparametric model: we are estimating a finite-dimensional parameter θ_0 and consider the infinite-dimensional η that describes the distribution of the data to be a nuisance parameter. Consider some (smooth) parametric submodel, so that the distribution is described by a finite-dimensional parameter. The Cramer-Rao lower bound guarantees a lower bound on the variance of any regular estimator in this parametric submodel. Now consider a semiparametric estimator that is regular in every parametric submodel. The variance of this estimator must be at least as large as the supremum of the lower bounds in all parametric submodels. This supremum is called the semiparametric efficiency bound (SPEB). More information about regularity and the semiparametric efficiency bound can be found in Bickel et al. (1993), Newey (1990), and van der Vaart (2000, Chapter 25).

For many econometric models with a random sample, we can use the methods for calculating the SPEB proposed in Newey (Newey1990) and Severini and Tripathi (2001). For the following theorem, the result for conditional moment restrictions for singular covariance matrices in Newey (2001) that extends a result in Chamberlain (1987) is important. The result shows that the optimal GMM estimator $\hat{\theta}_{W^*(n)}$ is asymptotically efficient for θ_0 among all regular semiparametric estimators.

THEOREM 7.2. Under assumptions MI, IID1, FULL-RANK, and FINITE-Ω_j, the semiparametric efficiency bound for θ_0 is $\text{SPEB}(\theta_0) = B^{-1}$, where B is as in (7.3).

REMARK 7.5. For specific examples, it may be reasonable to assume that $\Omega_j = \Omega$ and $D_j = D$ for each j. In that case, the expression for B simplifies to $B = D' \left(\sum_{j=1}^{J} p_j (S_j \Omega S_j)^+ \right) D$. This possibly lowers the SPEB, and our estimator may no longer be efficient.

7.4 Subsample estimation

In some situations θ_0 can be estimated using each subsample. An example is instrumental variable estimation with, for each pattern, more instruments than endogenous variables. We show that an optimal linear combination of the optimal GMM estimators for each subsample is asymptotically efficient. We study this estimator to gain more intuition for the semiparametric efficiency bound, and because it can be implemented using only the full-data estimation routine. Moreover, this estimator can be extended, without modification, to generalized empirical likelihood estimation. In Section 7.5, we generalize this approach to an optimal inverse probability weighting estimator for estimation under an assumption weaker than MI that allows for selection on observables.

Assume that θ_0 can be estimated using each subsample separately. Then the following subsample GMM estimator for θ_0 is defined for each missing data pattern j:

$$\hat{\theta}_{n,j} = \text{argmin}_{\theta\in\Theta}\, h_{n,j}(\theta)' W_j^*(n) h_{n,j}(\theta),$$

where $W_j^*(n)$ converges to the optimal weighting matrix $W_j^* = (S_j\Omega_j S_j)^+$. We look at matrix-weighted sums of these subsample GMM estimators. In particular, we are interested in the matrix weights that minimize the asymptotic variance of the sum. To find these, we need the limiting distribution of the subsample GMM estimators. Assume a standard GMM setting as in Section 7.3.1. Then, as $n \to \infty$,

$$\sqrt{n_j}(\hat{\theta}_{n,j} - \theta_0) \xrightarrow{d} N\left(0, \left(D_j'(S_j\Omega_j S_j)^+ D_j\right)^{-1}\right). \tag{7.5}$$

A matrix-weighted sum is the matrix equivalent of a weighted average. The weights are $p \times p$ matrices that are subsample specific, $(A_j(n),\ n \in \mathbb{N})$. An estimator that is a matrix-weighted sum is characterized by a $J-$tuple $A(n) =$

$(A_1(n), \ldots, A_J(n))$ that collects the matrix weights. We denote the matrix-weighted sum with matrix weights $A(n)$ by $\hat{\theta}_{A(n)}$, and define

$$\hat{\theta}_{A(n)} = \sum_{j=1}^{J} A_j(n)\hat{\theta}_{n,j}.$$

Assuming $\sum_{j=1}^{J} A_j = I_p$, the estimator is consistent. Since we have assumed a random sample, the subsample GMM estimators are uncorrelated, so that the asymptotic variance of matrix-weighted sum $\hat{\theta}_{A(n)}$ is given by

$$\lim_{n\to\infty} \text{var}(\sqrt{n}\hat{\theta}_{A(n)}) = \sum_{j=1}^{J} \frac{1}{p_j} A_j \left(D_j'(S_j\Omega_j S_j)^+ D_j\right)^{-1} A_j',$$

which uses the asymptotic variance of the subsample GMM estimators in (7.5). From the following theorem, we can see that the choice of weight matrix A_j^*,

$$A_j^* = B^{-1}\, p_j D_j'(S_j\Omega_j S_j)D_j,$$

leads to an efficient estimator $\hat{\theta}_n^* = \hat{\theta}_{A^*(n)}$. The asymptotic variance is

$$B^{-1} = \left(\sum_{j=1}^{J} p_j D_j'(S_j\Omega_j S_j)^+ D_j\right)^{-1}.$$

The theorem below shows that this is a lower bound for the asymptotic variance of any matrix-weighted sum.

THEOREM 7.3. For each $j = 1, \ldots, J$, let A_j be a $p \times p$ matrix such that $\sum_{j=1}^{J} A_j = I_p$. Then

$$\sum_{j=1}^{J} \frac{1}{p_j} A_j \left(D_j'(S_j\Omega_j S_j)^+ D_j\right)^{-1} A_j' - B^{-1}$$

is positive semidefinite.

Therefore, the estimator is the optimal linear combination of the optimal GMM estimators for each subsample. As such, it does not contain any additional nonlinear or nonparametric ingredients, which suggests that the higher-order asymptotic properties and small-sample performance of the efficient estimator under MI are of the same order as those of the full-data optimal GMM estimator.

REMARK 7.6. The discussion in this section suggests the following procedure to obtain an efficient estimator: (1) estimate $B = \sum_{j=1}^{J} p_j D_j' (S_j \Omega_j S_j)^{+} D_j$; (2) estimate $A_{j,n}^{*} = B^{-1}\, p_j D_j' \left(S_j \Omega_j S_j\right)^{+} D_j$; (3) set $\hat{\theta}_{A^*(n)} = \sum_{j=1}^{J} A_{j,n}^{*} \hat{\theta}_{j,n}$.

REMARK 7.7. The results in this section can be used to optimally combine estimators obtained using any estimation method, provided that the data used for different estimators is independent. For example, the results can be applied to generalized empirical likelihood estimation. Another example is a combination of estimators applied to different data sets.

7.5 Inverse probability weighting

In the previous section we derived an optimal estimator under MI and IID1. For some applications, the MI assumption is too strong. In this section, we introduce a weaker assumption about the missing data mechanism, CMI, that allows the missing data indicator to depend on some observed random variables. We generalize the inverse probability weighting (IPW) estimator to a class of estimators that are consistent under CMI. Then, we use techniques from Sections 7.3 and 7.4 to derive the efficient IPW estimator.

7.5.1 Missing at random

For many situations, both MCAR and MI are too strong. A significantly weaker assumption that can be used is missing at random, MAR. Organize the data into two groups, (X, Z), where $X \in \mathbb{R}^d$, $Z \in \mathbb{R}^{d_z}$. The random vector X enters

the moment function, but the random vector Z does not; it is a vector of auxiliary variables. The missing data pattern for X is captured by $R^x \in \mathbb{R}^{d\times d}$, a random matrix that takes values $\{S_1^x, \ldots, S_{J_x}^x\}$. The following assumption is a typical version of MAR, although different versions are possible:

ASSUMPTION 7.11 (MAR). For each pattern j, $X \perp R^x \mid Z$.

ASSUMPTION 7.12 (IID2). $(R_i^x, R_i^x X_i, Z_i,\ 1 \leqslant i \leqslant n)$ is a random sample of size n from $(R^x, R^x X, Z)$.

The MAR assumption allows the process that generates the missing data to depend on that data. It requires that there exists an auxiliary random vector Z that is always observed and that removes the dependence between R^x and X. This is a significantly weaker assumption than MCAR, especially when many relevant variables are included in Z.

We will formulate an assumption that relaxes MAR in the way that MI relaxes MCAR. As in the MCAR case, the missing data indicator R^x implies a missing data indicator R that describes which components of $h(X,\theta)$ can be evaluated when $R^x X$ is observed instead of X. There are J such patterns for h, denoted $\{S_1, \ldots, S_J\}$.

Consider pattern j, and let r_j be an indicator function that equals 1 if and only if the missing data follow pattern j. Let $V_j \in \mathbb{R}^{d_j}$ be a random vector that consists of a subset of the components of (X, Z). We assume that there exists a function p_j that determines the probability of observing pattern j : $p_j(V_j) = \mathrm{P}(r_j = 1 \mid V_j)$. Let $V = \cup_{j=1}^J V_j$.

ASSUMPTION 7.13 (CMI). (i) $\mathrm{E}\left(h(X,\theta_0) \mid r_j,\ V_j\right) = \mathrm{E}\left(h(X,\theta_0) \mid V_j\right)$; (ii) $p_j(V_j)$ is observed if $r_j = 1$; (iii) $\mathrm{P}(r_j = 1|V) = \mathrm{P}(r_j = 1|V_j)$; (iv) there exists $\delta > 0$ such that $p_j(V_j) \geqslant \delta$ for each V_j.

The first assumption captures the essence of MAR, and assumptions (ii)–(iv) are necessary for the construction of an inverse probability weighted estimator in Section 7.5.2. We are not interested in the function $p_j(V_j)$ and assume that the function is known or can be $\sqrt{n_j}-$consistently estimated, which under CMI is not very restrictive given the results in Hirano et al.

(2003). Notice that elements of X can be included in V_j if they are observed whenever $r_j = 1$. Also, missing data indicators r_k, $k \neq j$, can be included, provided the resulting p_j obeys CMI (iv).

7.5.2 Optimal IPW

A standard tool for missing data with a binary missing data pattern that satisfies MAR is inverse probability weighting (IPW); see for example Wooldridge (2007). In this section we consider a generalization of IPW estimators to the case of general missing data patterns. The assumption of CMI ensures the consistency of such an IPW estimator. First, note that we can rewrite $R = \sum_{j=1}^{J} r_j S_j$. If we have a function $h(X, \theta_0)$ for which $\mathrm{E}(h(X, \theta_0)) = 0$ then, in general, $\mathrm{E}(Rh(X, \theta_0)) \neq 0$. Now let $\tilde{R}(V) = \sum_{j=1}^{J} \frac{r_j}{p_j(V_j)} S_j$. Then

$$\mathrm{E}\left(\tilde{R}(V)\middle| V\right) = \sum_{j=1} \frac{\mathrm{E}(r_j|V_j)}{p_j(V_j)} S_j = \sum_{j=1}^{J} S_j.$$

and, using iterated expectations, $\mathrm{E}\left(\tilde{R}(V)h(X, \theta_0)\right) = 0$.

This motivates the use of the adjusted subsample moment $\tilde{h}_{n,j}$,

$$\tilde{h}_{n,j} = \frac{1}{n_j} \sum_{i \in G_j} \frac{1}{p_j(V_j)} R_i h(X_i, \theta_0).$$

An IPW version of the complete-case estimator minimizes $\tilde{h}'_{n,1} W^*_{\mathrm{cc}}(n) \tilde{h}_{n,1}$, and an IPW version of the available-case estimator minimizes

$$\left(\sum_{j=1}^{J} \tilde{h}_{n,j}\right)' W^*_{\mathrm{ac}}(n) \left(\sum_{j=1}^{J} \tilde{h}_{n,j}\right),$$

where the respective W^* can be chosen optimally.

This suggests an extension of the method in Section 7.4. Assume that θ_0 can be estimated using each subsample separately. Furthermore, the assumptions for asymptotic normality of the optimal GMM estimator and CMI hold.

Then, the parameter θ_0 is identifiable within subsample G_j. Denote the optimal subsample IPW estimator $\hat{\tilde{\theta}}_{n,j}$:

$$\hat{\tilde{\theta}}_{n,j} = \text{argmin}_{\theta \in \Theta} \tilde{h}_{n,j}(\theta)' W_j^*(n) \tilde{h}_{n,j}(\theta), \tag{7.6}$$

with W^* equal to the optimal weighting matrix for this problem. The limiting distribution of $\hat{\tilde{\theta}}_{n,j}$ is that of a standard GMM estimator: as $n_j \to \infty$,

$$\sqrt{n_j}(\hat{\tilde{\theta}}_{n,j} - \theta_0) \xrightarrow{d} N\left(0, \Lambda_j\right).$$

We do not impose any structure on Λ_j, since we have not specified whether the function is known, or whether a parametric or nonparametric estimator was used.

Analogously to Section 7.4, we introduce the class of estimators

$$\hat{\tilde{\theta}}_{A(n)} = \sum_{j=1}^{J} A_j(n) \hat{\theta}_{n,j}, \tag{7.7}$$

for any $J-$tuple of $p \times p$ matrices $A(n) = (A_1(n), \ldots, A_J(n))$ that satisfies $\sum_{j=1}^{J} A_j(n) = I_p$. For each sequence $A(n)$ that converges to some A, the asymptotic variance is given by

$$\lim_{n \to \infty} \text{var}(\sqrt{n} \hat{\tilde{\theta}}_{A(n)}) = \sum_{j=1}^{J} \frac{1}{p_j} A_j \Gamma_j A_j'.$$

A straightforward modification of Theorem 7.3 shows that the lower bound on the asymptotic variance for any estimator in the class of matrix-weighted sums is given by

$$\tilde{B}^{-1} = \left(\sum_{j=1}^{J} p_j \Gamma_j \right)^{-1}.$$

Setting $A_j^* = \tilde{B}^{-1} \, p_j \Gamma_j$ achieves that bound.

7.6 Examples

This section contains four examples that illustrate the methods in this chapter and demonstrate the efficiency gains with respect to a complete-case and an available-case analysis. The first example concerns a multivariate mean estimation problem that corresponds to a two-period panel data model with attrition. In the second example, we discuss an instrumental variable model where the instruments are partially observed. The third example is the estimator proposed by Arellano and Bond (1991) for dynamic panel data models. In the fourth example, we use our results to optimally design a data set to measure the change in consumer confidence when nonresponse is expected. The derivations are available upon request.

7.6.1 Attrition in two periods

We study a two-period panel data model with attrition as an example of multivariate mean estimation with missing data. We present analytical results for the asymptotic variance of the estimators.

A health club is interested in measuring the change in the weight of new members after they join. New members are weighed upon registration, and a random sample of new members is selected to come back for a reweighing after six months. Let $X_{i,1}$ be the weight of member i upon registration and let $X_{i,2}$ be the weight of that member after six months.

An error component model can be used to model $X_i = (X_{i,1}, X_{i,2})$: $X_{i,t} = \mu_t + \alpha_i + \epsilon_{it},\ t = 1, 2$, where $\mathrm{E}(\alpha_i) = 0,\ \mathrm{var}(\alpha_i) = \sigma_a^2$ and $\mathrm{E}(\epsilon_{it}) = 0,\ \mathrm{var}(\epsilon_{it}) = \sigma_e^2$ for each $t = 1, 2$. We normalize $\sigma_a^2 + \sigma_e^2$ and denote $\rho = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$. As a result, $\mathrm{E}(X_i) = (\mu_1, \mu_2)$ and $\Omega = \mathrm{var}(X_i) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

There are two missing data patterns, corresponding to two groups. For an observation i in the first group we observe both $X_{i,1}$ and $X_{i,2}$. For an observation in group 2 we observe only $X_{i,1}$. In other words, $d = 2,\ q = 2,\ J = 2, S_1 = I_2$, and $S_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. Assuming that all members who are called for a reweighing show up, the health center has full control over the

Estimator	avar $(\hat{\mu}_2)$	avar $(\hat{\mu}_2 - \hat{\mu}_1)$
full data	1	$2\ (1-\rho)$
complete case	$1/p_1$	$2\ (1-\rho)/p_1$
available case	$1/p_1$	$(1-2\rho) + 1/p_1$
optimal	$1/p_1(1-\rho^2(1-p_1))$	$(1-2\rho) + 1/p_1\ (1-\rho^2(1-p_1))$

Table 7.2: Comparison of asymptotic variances.

randomization mechanism, so we assume MI and $\Omega_1 = \Omega_2 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Finally, define $p_1 = \mathrm{P}\,(R = S_1)$.

The estimation is focused on μ_2 and $(\mu_2 - \mu_1)$ and based on the moment conditions $E(h(X,\mu)) = \mathrm{E}(X - \mu) = 0$. We consider four estimators. The first is the full-data estimator, which equals the sample mean using all n observations. This estimator is not feasible because it uses observations that are missing. We include this estimator to quantify the amount of information that is lost because of the missing data. The second estimator is the complete-case estimator and uses only the complete observations in group 1. The third estimator is the available-case estimator. This estimator uses the maximum number of observations per component: $n_1 + n_2$ for μ_1 and n_1 for μ_2. Finally, we consider the optimal sample mean.

The asymptotic variances of the estimators in this example for $\hat{\mu}_2$ and $(\hat{\mu}_2 - \hat{\mu}_1)$ are given in Table 7.2. In Figures 7.2 and 7.3 we compare the variances as a function of ρ.

The key element of this example is the individual effect, which introduces correlation between the components of X_i. The optimal estimator efficiently exploits this correlation. An interesting finding is that including observations for members who are observed only upon registration increases the precision for the average weight after six months and for the average change in weight.

The first column of Table 7.2 and Figure 7.2 show that, for estimating μ_2, the complete-case and the available-case estimators do not recover any of the information that is lost because of the missing data, even when the components are highly correlated. The optimal estimator efficient exploits the correlation. As the individual effect becomes more important, the performance of the

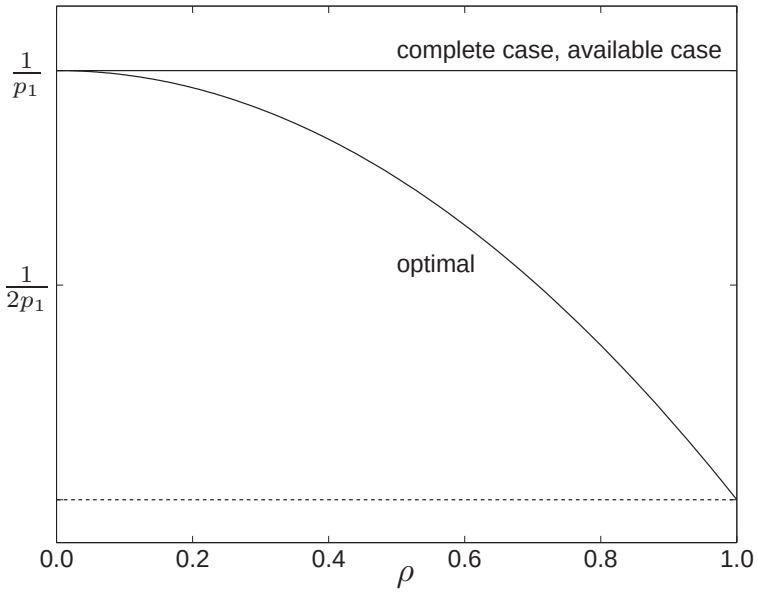


Figure 7.2: Asymptotic variances of $\hat{\mu}_2$ as a function of ρ.

optimal estimator relative to the full-data estimator improves. In particular, if $\rho = 1$, observing $X_{i,2}$ does not give any additional information, and the optimal estimator is as efficient as the full-data estimator.

The second column of Table 7.2 and Figure 7.3 describe the relative performance of the estimator of $\mu_2 - \mu_1$. All estimators benefit from the correlation between X_{i1} and X_{i2}. In the absence of correlation, the optimal estimator coincides with the available-case estimator. If the components are perfectly correlated, both the optimal estimator and the complete-case estimator retrieve all the information.

To understand why the relative performance of the complete-case and the available-case estimators depends on the correlation, consider that the complete-case estimator corresponds to first calculating $X_{i,2} - X_{i,1}$ and then averaging, while the available-case estimator averages the $X_{i,1}$ and the $X_{i,2}$ and then takes the difference. For the complete-case estimator the individual effects drop out, so that high values of σ_a^2 (ρ) are not reflected in the variance of the estimator. For the variance of the available-case estimator, σ_a^2 does play a role, because this estimator includes observations for which only one period is available. An increase in σ_α^2 therefore increases the variance of the

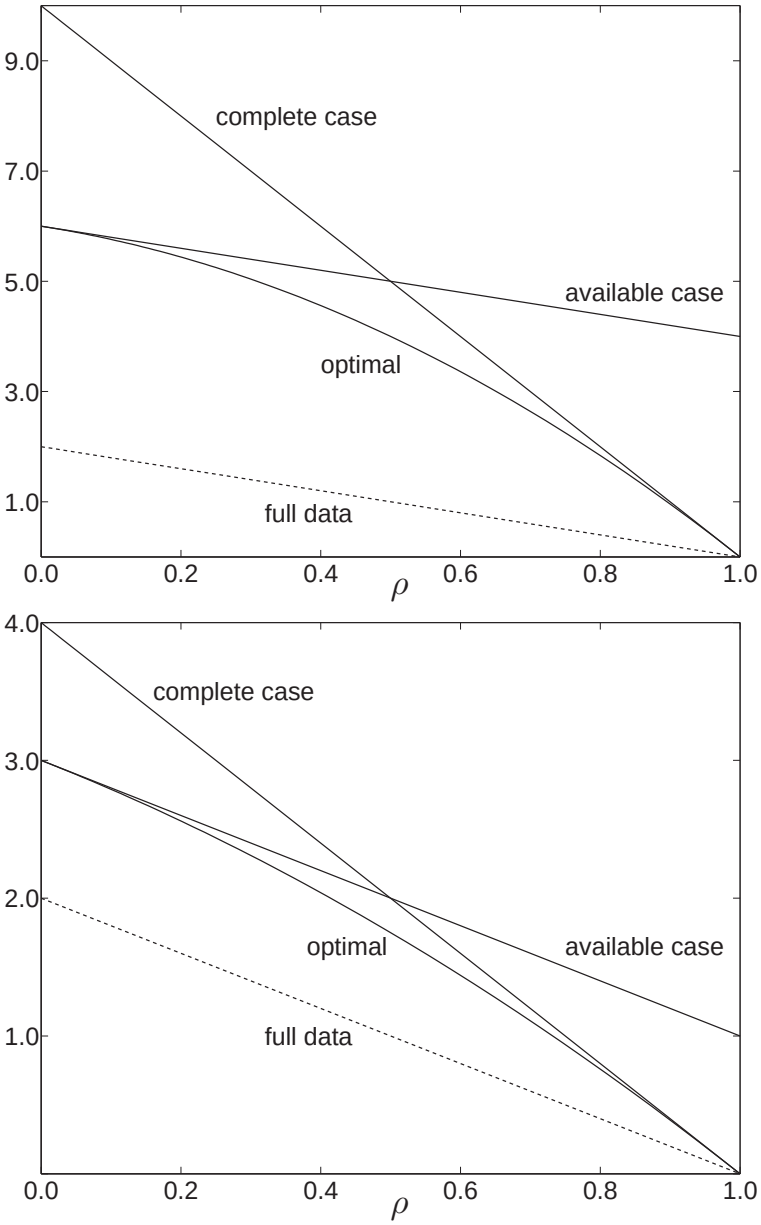


Figure 7.3: Asymptotic variances of $\hat{\mu}_2 - \hat{\mu}_1$ as a function of ρ. Top panel: $p_1 = 0.2$. Bottom panel: $p_1 = 0.5$.

available-case estimator.

7.6.2 Instrumental variables

We study a simple linear instrumental variable model where the dependent and explanatory variables are always observed, but instruments can be incomplete. We consider the linear case with one explanatory variable and two instruments. Either instrument can be missing for a subsample. The approach is easily generalized to multiple explanatory variables, multiple instruments, and nonlinear models. The setup in this section has the advantage that it allows us to derive analytical results. The problem of partially missing instruments is common; a recent example can be found in Angrist et al. (2006).

The dependent variable y is linearly related to an explanatory variable x, $y = \beta x + \epsilon$. Two instruments, w_1 and w_2, are available, which motivates the following unconditional moment conditions to estimate β:

$$0 = \mathrm{E}\left(\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}(y - \theta_0 x)\right) = \mathrm{E}\left(\begin{pmatrix} w_1\epsilon \\ w_2\epsilon \end{pmatrix}\right).$$

We assume that the dependent variable and the explanatory variable are always observed. There are three groups of observations, $J_x = 3$. For the first group we observe both instruments. For the second group we observe only w_1, and for the third group we observe only w_2. As a result, $J = 3$ and

$$S_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ S_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},\ S_3 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

We assume that the instruments are similar: they are equally likely to be missing, $p_2 = p_3 = (1 - p_1)/2$, they have the same correlation with the explanatory variable, $\mathrm{E}(w_1 x) = \mathrm{E}(w_2 x) = \lambda$, and they are both standardized so that $\mathrm{E}(w_j) = 0,\ j = 1, 2$ and $\mathrm{E}(w_j^2) = 1,\ j = 1, 2$. The instruments have correlation $\rho = \mathrm{cov}(w_1, w_2)$.

We assume that the variance matrices are the same for all groups:

$$\Omega_1 = \Omega_2 = \Omega_3 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

where the form of Ω could result from the additional assumptions $\mathrm{E}(w_j^2\epsilon^2) = \mathrm{E}(w_j^2)\,\mathrm{E}(\epsilon^2) = 1,\ j = 1, 2$. Furthermore, we normalize the variance of the explanatory variable, $\mathrm{var}(x) = 1$. Since $\mathrm{var}(x, w_1, w_2)$ must be semidefinite, we have

$$\mathrm{var}(x, w_1, w_2) = \begin{pmatrix} 1 & \lambda & \lambda \\ \lambda & 1 & \rho \\ \lambda & \rho & 1 \end{pmatrix},$$

$$|\mathrm{var}(x, w_1, w_2)| = (-1)\rho^2 + (2\lambda^2)\rho + (1 - 2\lambda^2),$$

and it follows that $\rho \geqslant 2\lambda^2 - 1$. We fix $\lambda = \frac{1}{\sqrt{2}}$ so that the lower bound for ρ is 0. This assumption does not affect the relative efficiency of the estimators.

We consider five estimators. The first four (full data, complete case, available case, and optimal) have been discussed in the text and in Example 7.6.1. The fifth, which we call the complete-moment estimator, uses one moment only. Because the instruments are similar, the two complete-moment estimators have the same asymptotic variance.

In Figure 7.4 we plot the asymptotic variance of our estimators as a function of ρ for $p_1 = 0.5$. The key aspect of this example is that the two instruments act as similar sources of information for estimating β. Therefore, as the correlation between w_1 and w_2 increases, we expect two effects. First, the total amount of information for β decreases, so we expect all estimators to be worse. Secondly, the amount of information on the instrument that is missing increases. Since the optimal estimator is constructed such that it efficiently exploits the correlation between the components of the moment conditions, we expect the relative performance of the optimal estimator to increase.

The optimal estimator is efficient among the feasible estimators. Except for $\rho = 0$, it outperforms the available-case estimator. As ρ increases, the

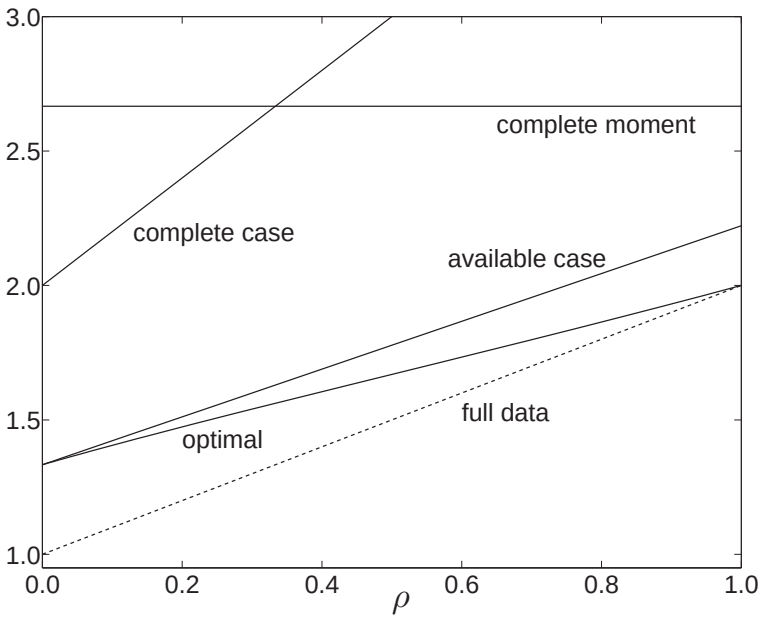


Figure 7.4: Asymptotic variance for various estimators of β as a function of ρ, $p_1 = 0.5$.

relative performance of the optimal estimator with respect to the available estimator increases: the available-case estimator uses all the available data but does not efficiently use the correlation between the instruments. As ρ approaches 1, the optimal sample mean is able to recover all the information. The complete-case and complete-moment estimators are always outperformed by the available-case estimator and the optimal sample mean.

7.6.3 Dynamic panel data

The goal of this setting is to demonstrate the performance of our method in a more complicated model and to provide an example where the variance matrix is not known. In particular, we look at a dynamic panel data model, and use continuous updating GMM to estimate it.

The parameter of interest ρ describes the relationship between current and lagged values of a random variable $y_{i,t}$, $y_{i,t} = \alpha_i + \rho y_{i,t-1} + \epsilon_{i,t}$, $2 \leqslant t \leqslant T$. We assume that $\mathrm{E}(\alpha_i) = 0$, $\mathrm{var}(\alpha_i) = \sigma_a^2$, and $\mathrm{E}(\epsilon_{it}) = 0$, $\mathrm{var}(\epsilon_{it}) = \sigma_e^2$, and $\mathrm{E}(\epsilon_{i,t}\epsilon_{i,s}) = 0$ whenever $s \neq t$. Arellano and Bond (1991) propose an estimator

that is widely used: the optimal GMM estimator based on the $(T-2)(T-1)/2$ moment conditions $\mathrm{E}(y_{i,t-s}\Delta\epsilon_{i,t}) = 0,\ t \geqslant 3, s \geqslant 2$.

For any observation i, if $y_{i,t}$ is not observed, then several components of the moment function are not observed. For an example with $T = 5$, see Table 7.1 in the introduction. For the purposes of this simulation, we consider the case $T = 9$, which corresponds to the example in Blundell and Bond (1998). This gives 28 moment conditions for 1 parameter. If any of the $y_{i,t}$ are missing, the moment function is incompletely observed: if $y_{i,1}$ is not observed, 7 components of the moment function are not observed; if $y_{i,4}$ is not observed, 12 components of the moment function are not observed.

We perform a Monte Carlo analysis to compare the relative performance of the estimator introduced in this chapter to the full-data, complete-case, and available-case estimators. We do not assume the variance matrix to be known, and use a continuous updating version of the Arellano-Bond estimator to estimate ρ. When estimating the variance matrix, we assume that $\Omega_j = \Omega$ for each j.

We normalize $\sigma_\epsilon^2 = 1$. We consider different values for the variance of the individual effect $\sigma_\alpha^2 \in \{0.1, 1\}$ and the parameter of interest $\rho \in \{0.1, 0.2, 0.5, 0.8\}$. We set $n = 10000$ and perform $s = 1000$ simulations per parameter combination. There are 10 missing data patterns. Patterns $j = 1, \ldots, 9$ have $y_{i,j}$ missing and the other variables observed. Pattern 10 corresponds to the subsample with all variables observed. This missing data pattern is determined by a parameter p such that $p = \mathrm{P}(R = S_j)$ for each $j = 1, \ldots, 9$, and $\mathrm{P}(R = S_{10}) = 1 - 9p$. We consider $p \in \{0.02, 0.06\}$ so that 82% (respectively 46%) of the observations are complete.

Table 7.3 reports the variance of the complete-case, available-case, and optimal estimator divided by the variance of the full-data estimator. The complete-case estimator is always worse than the available-case estimator, except for $(\sigma_\alpha^2, \rho, p) = (1, 0.8, 0.02)$. The optimal estimator always outperforms the other two estimators. In contrast to the case where the Ω_j are known, this is not true by construction. The optimal estimator seems to gain more when p is larger. For some parameter configurations, the efficiency gain is substantial.

σ_α^2	ρ	p	cc	ac	opt
0.1	0.1	0.02	1.19	1.12	1.08
		0.06	2.29	1.46	1.41
	0.2	0.02	1.29	1.23	1.18
		0.06	2.37	1.34	1.27
	0.5	0.02	1.82	1.77	1.69
		0.06	3.35	2.50	2.25
	0.8	0.02	8.61	8.11	7.74
		0.06	15.95	11.76	10.45
1	0.1	0.02	1.71	1.47	1.46
		0.06	3.04	1.89	1.84
	0.2	0.02	1.91	1.70	1.68
		0.06	3.75	2.35	2.21
	0.5	0.02	5.10	4.75	4.59
		0.06	8.61	5.85	5.33
	0.8	0.02	2.04	2.20	1.92
		0.06	3.47	3.30	2.62

Table 7.3: Relative variance of the complete-case (cc), available-case (ac), and optimal (opt) estimator in a Monte Carlo study of a continuous updating Arellano-Bond estimator, with $n = 10000$, $s = 1000$, and $T = 9$. The missing data patterns are described in the text.

7.6.4 Panel design

We have considered optimal estimation for given missing data patterns. This analysis is useful for many applications in economics, where the researcher has no control over the data-collection process. For the data collector the relative performance of estimators under different missing data patterns is of importance. Assuming that the researcher uses efficient methods to deal with missing data, what is the best way to collect the data? We discuss data collection for a variable that varies over individuals and over time. We are interested in estimating the change in the population average of the variable over time. We consider three ways to collect the data: repeated cross-sections, a panel, and a rotating panel.

A researcher wants to measure the change in consumer confidence over a period of three years. Denote the confidence of consumer i at time t by

$X_{i,t}$, where $1 \leqslant t \leqslant 3$, which can be modeled using error components: $X_{i,t} = \alpha_i + \mu_t + \epsilon_{i,t}$. The level of consumer confidence at time t is μ_t. Some consumers may have, across all periods, a more optimistic or pessimistic outlook on the economy, and this is captured by α_i, $\mathrm{E}(\alpha_i) = 0$, and $\mathrm{var}(\alpha_i) = \sigma_a^2$. The idiosyncratic error term $\epsilon_{i,t}$ captures random errors in the elicitation process, and we assume that $\mathrm{E}(\epsilon_{i,t}) = 0$ and $\mathrm{var}(\epsilon_{i,t}) = \sigma_e^2$. It follows that

$$\mathrm{var}(X_{i,t}) = (\sigma_a^2 + \sigma_e^2) \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix},$$

where $\rho = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$. The level of consumer confidence does not have an interpretation, so we normalize $\sigma_a^2 + \sigma_e^2 = 1$. The parameters of interest are the changes in consumer confidence, $\mathrm{E}(X_{i,t} - X_{i,t-1}) = \mu_t - \mu_{t-1} = \delta_{t-1}$, for $t = 2, 3$.

The researcher has a budget of $\$M$. Surveying a person once costs \$1, so the researcher can obtain at most M consumer confidence measurements. She considers three ways of collecting the data. The first is a repeated cross-section: for each period, survey a random sample of $M/3$ consumers from the population. The second is a panel: randomly select $M/3$ consumers and survey them in each period. The third is a rotating panel: randomly select $M/4$ consumers to survey in periods 1 and 2, and randomly sample $M/4$ consumers for periods 2 and 3. All these methods exhaust the research budget.

Not all the surveys are completed, which leads to missing data. The missing data mechanism is assumed to be MI. The probability that a consumer does not respond, or stops responding, is p. The research budget allocated to this consumer is lost. Once the data are collected, the researcher will use the methods in this chapter to estimate δ_1 and δ_2 optimally. Figures 7.5 and 7.6 show the asymptotic variance of $\hat{\delta}_1$ and $\hat{\delta}_2$ for each of the approaches for $p = 0.1$ and $p = 0.5$ respectively.

The relative performance of the cross-section method increases as the probability of nonresponse increases: a panel member is lost forever, so the effect of nonresponse for the (rotating) panel is stronger than for the cross-section.

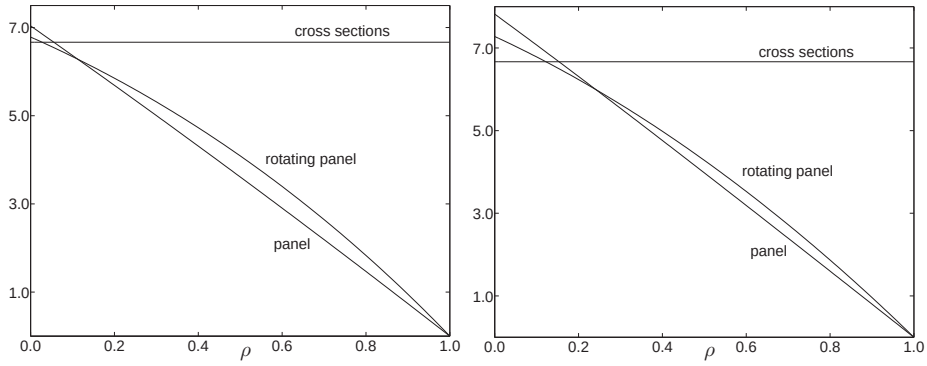


Figure 7.5: Asymptotic variances of optimal estimators of the change in consumer confidence using different data collection methods; $p = 0.1$. Left panel: δ_1. Right panel: δ_2.

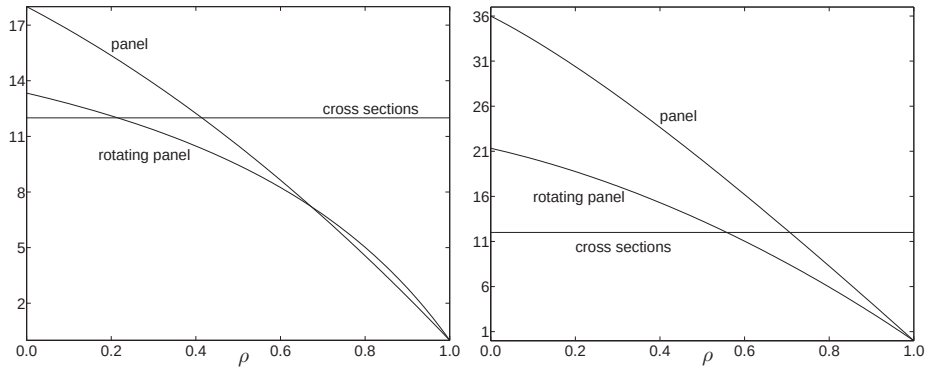


Figure 7.6: Asymptotic variances of optimal estimators of the change in consumer confidence using different data collection methods; $p = 0.5$. Left panel: δ_1. Right panel: δ_2.

As ρ increases, the relative performance of the cross-section method decreases, since there is no information available on the missing data, whereas the panel methods can extract some information through the individual effect. The variance of the panel methods is similar for $p = 0.1$, but the rotating panel leads to more substantially more efficient estimators for $p = 0.5$.

7.7 Conclusion

This chapter considered efficient GMM estimation from a random sample of complete and incomplete observations. We derived the semiparametric efficiency bound under an assumption that is weaker than missing completely at random. We introduced an efficient estimator by assigning observations to subsamples on the basis of their missing data pattern. This approach allows us to extend the estimator to a setting where selection is on unobservables. Examples demonstrated the flexibility of the approach and the efficiency gains that can be obtained over standard approaches.

Some aspects of the chapter could be further investigated. First, the framework that we constructed to deal with a general missing data pattern suggests some tests for sample selection. In particular, if the parameter is identifiable in each subsample, a test of equality of the subsample estimators can be used to detect sample selection. Second, the mathematical result underlying Section 7.4 may be of independent interest. We will explore extensions and further applications in future work.

7.A Proofs

Proof of Theorem 7.1. Abbreviate Newey and McFadden (1994) to NM94. We are going to construct a function Q_0 such that conditions (i)-(iv) in (NM94, Theorem 2.1) are satisfied with respect to Q_0 and Q_n. We defined Q_n in 7.2.

Construction of Q_0. Note that

$$\begin{aligned}
\mathrm{E}\left(Rh(X,\theta)\middle|\, R = S_j\right) &= S_j \int h(x,\theta) f_{x|j}(x) dx \\
&= \frac{1}{p_j} S_j \int h(x,\theta) f_{x,r}(x,r) dx \\
&= \frac{1}{p_j} S_j \int h(x,\theta) f_j(x) f_x(x) dx \\
&= \frac{1}{p_j} S_j \,\mathrm{E}\left(h(x,\theta) f_j(x)\right).
\end{aligned}$$

Consider the function

$$k(x,\theta) = \begin{pmatrix} \frac{1}{p_1} S_1 h(x,\theta) f_1(x) \\ \vdots \\ \frac{1}{p_J} S_J h(x,\theta) f_J(x) \end{pmatrix}.$$

Form the blockdiagonal matrix W_n from the blocks $(W_{1,n}, \ldots, W_{j,n})$, and let $W_n \to W$. Define $Q_0(\theta) = k(\theta)' W k(\theta)$. This function can be seen as a GMM criterion function for the Jq moment conditions implied by the conditional moment restrictions $\mathrm{E}(Rh(X,\theta_0)|R) = 0$.

Identification (and compactness) - i and ii. Because of MI, IDENTIFICATION, and REGULARITY(iii), $Q_0(\theta)$ has a unique minimum at θ_0. Therefore, condition (i) for (NM94, Theorem 2.1) is satisfied. Condition (ii) is automatically satisfied by REGULARITY(i).

Continuity - iii. Continuity of k follows immediately from the continuity of h and the requirement that f_j does not depend on θ. This implies that $\frac{1}{p_J} S_j h(x,\theta) f_j(x)$ is continuous in θ if h is.

Uniform convergence - iv. The sample average for subsample j, $\tilde{h}_j(\theta)$, converges uniformly to the j-th conditional expectation. First, we show why

the subsample average would converge to the conditional expectation if the inner function were bounded. Then we show that convergence is uniform.

Consider the sample average $\frac{1}{n}\sum_i 1\{R_i = S_j\}R_i h(X_i,\theta)$. By the law of large numbers, this converges to

$$\begin{aligned}\mathrm{E}(1\{R = S_j\}Rh(X,\theta)) &= \sum_k p_k \,\mathrm{E}(1\{R = S_j\}Rh(X,\theta)|\, R = S_k) \\ &= p_j S_j \,\mathrm{E}(h(X,\theta)|\, R = S_j)\end{aligned}$$

This implies that the subsample average $\tilde{h}_j \to S_j \,\mathrm{E}(h(X,\theta)|\, R = S_j)$. This shows convergence for each component of k, and therefore for k, and therefore, with condition 6, for Q.

For each component j of the function k, equals $\frac{1}{p_j}S_j h(x,\theta)f_j(x)$. Since $f_j \in [0,1]$ and $p_j \in (0,1]$, the boundedness of h translates to boundedness of k. Continuity of k follows from continuity of h. Therefore, convergence is uniform. See (NM94, Lemma 2.4).

Conditions (i)-(iv) of NM94 are satisfied, and hence $\hat{\theta}_n \to \theta_0$. □

Proof of Theorem 7.2. Each observation provides two random objects that we can use for estimation: a missing moment indicator R_i and the observed elements of the moment function $R_i h(X_i,\dot{})$. The moment conditions are provided by MI, which states that $\mathrm{E}(Rh(X,\theta_0)\ |R)) = 0$. Furthermore, we have that $\mathrm{E}(R) = \sum_{j=1}^{J} p_j S_j$. Under the typical MCAR assumption, we have more information about R, which we can exploit as additional moment conditions, see Graham (2010). However, MI does not provide conditional moment conditions of R on X, or some function of X.

Therefore, the model implies the following moment restrictions on our data: (i) $\mathrm{E}(R) = \sum_{j=1}^{J} p_j S_j$, and (ii) $\mathrm{E}(Rh(X,\theta_0)\ |R)) = 0$. First, we show that the unconditional moment restrictions (i) are not informative for θ_0. Then we derive SPEB(θ_0) using the conditional moment restrictions (ii).

First, denote

$$\mathrm{E}(Rh(X,\theta_0)\ |R)) = \mathrm{E}(\psi_1(R,X;\theta_0)\ |R)$$

and $\mathrm{E}(R - \sum_{j=1}^{J} p_j S_j) = \mathrm{E}(\rho_2(R;p))$, where $p = (p_1,\ldots,p_J)$. Since R has finite support, there exists a function $M(R)$ such that the unconditional moment restrictions

$$\mathrm{E}(M(R)\psi_1(R,X;\theta_0)) = \mathrm{E}(\rho_1(R,X;\theta_0)) = 0$$

contain the same information as $\mathrm{E}(\psi_1(R,X;\theta_0)\ |R) = 0$. Let $\beta_0 = (\theta_0, p)$. The asymptotic efficiency bound for β_0 based on the unconditional moment restrictions $\mathrm{E}(\rho(R,X;\beta_0)) = \begin{pmatrix} \rho_1(R,X;\theta_0) \\ \rho_2(R;p) \end{pmatrix} = 0$ is $\Lambda_0 = \left(D_0'\Sigma_0^{-1}D_0\right)^{-1}$, where $D_0 = \mathrm{E}\left(\frac{\partial\rho(R,X;\beta_0)}{\partial(\theta)}\right)$ and $\Sigma_0 = \mathrm{E}\left(\rho(R,X;\beta_0)\rho'(R,X;\beta_0)\right)$, following Chamberlain (1987). D_0 can be partitioned as $D_0 = \begin{pmatrix} \mathrm{E}(\frac{\partial\rho_1(\beta_0)}{\partial\theta}) & 0 \\ 0 & \mathrm{E}(\frac{\partial\rho_2(\beta_0)}{\partial p}) \end{pmatrix}$. The off-diagonal blocks of D_0 are zero, since θ_0 only features in ρ_1 and p only features in ρ_2. Therefore, the bound for θ_0 under $\mathrm{E}(\rho_1) = 0$ equals the bound for θ_0 under $\mathrm{E}(\rho) = 0$, and we conclude that ρ_2 is not informative for θ_0.

Next. we can find the semiparametric efficiency bound for θ_0 given the conditional moment conditions

$$\mathrm{E}(Rh(X,\theta_0)\ |R) = \mathrm{E}(\rho(R,X,\theta_0)\ |R) = 0$$

by applying the result in Newey (2001, Theorem 5.2) that extends Chamberlain (1987). Let $D_\rho(R) = \frac{\partial \mathrm{E}(\rho(X,R,\theta_0)|R)}{\partial \theta}$ and

$$\Sigma_\rho(R) = \mathrm{E}(\rho(X,R,\theta_0)\rho(X,R,\theta_0)'|R).$$

The semiparametric efficiency bound is equal to

$$\mathrm{SPEB}(\theta_0) = \left(\mathrm{E}\left(D_\rho(R)'\Sigma_\rho(R)^+ D_\rho(R)\right)\right)^{-1},$$

In our case, $D_\rho(S_j) = S_j D_j = S_j \,\mathrm{E}\left(\frac{\partial h(X,\theta_0)}{\partial \theta}\middle| R = S_j\right)$ and $\Sigma_\rho(S_j) = S_j \Omega_j S_j$. Then,

$$\begin{aligned}\mathrm{SPEB}(\theta_0) &= \left(\sum_{j=1}^{J} p_j D_j' S_j (S_j \Omega_j S_j)^+ S_j D_j\right)^{-1} \\ &= \left(\sum_{j=1}^{J} p_j D_j' (S_j \Omega_j S_j)^+ D_j\right)^{-1}.\end{aligned}$$

□

Proof of Theorem 7.3. Let $\Gamma_j = D_j'(S_j\Omega_j S_j)^+ D_j$. $\Gamma_j = \Gamma_j'$ and, because of IDENTIFICATION+, Γ_j is invertible. We need to show that, for any $J-$tuple of weighting matrices ($A_j \in \mathbb{R}^{p\times p},\ j = 1,\ldots,J$),

$$\sum_{j=1}^{J} \frac{1}{p_j} A_j \Gamma_j^{-1} A_j' - (\sum_{j=1}^{J} p_j \Gamma_j)^{-1}$$

is positive semidefinite. Let $K_1' = \left[1/\sqrt{p_1}A_1\Gamma_1^{-1/2} \quad \dots \quad 1/\sqrt{p_J}A_J\Gamma_J^{-1/2}\right]$, so that $K_1'K_1 = \sum_{j=1}^{J} \frac{1}{p_j} A_j\Gamma_j^{-1}A_j'$. Similarly, let

$$K_2' = \left[\sqrt{p_1}\Gamma_1^{1/2} \quad \dots \quad \sqrt{p_J}\Gamma_J^{1/2}\right],$$

so that $(K_2'K_2)^{-1} = (\sum_{j=1}^{J} p_j D_j'(S_j\Omega_j S_j)^+ D_j)^{-1}$.

Furthermore, $K_1'K_2 = \sum_{j=1}^{J} \sqrt{p_j}/\sqrt{p_j}A_j\Gamma_j^{-1/2}\Gamma_j^{1/2} = \sum_{j=1}^{J} A_j = I_p$. Then, by Magnus and Abadir (2005), $K_1'K_1 - (K_2'K_2)^{-1}$ is positive semidefinite, which completes the proof. □

Bibliography

Abramowitz, M., and I.A. Stegun, Eds. (1964), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, ninth Dover printing, tenth GPO printing, Dover, New York.

Abadir, K., and J.R. Magnus (2005), *Matrix Algebra*, Cambridge University Press, Cambridge.

Abowd, J. M., B. Crépon, and F. Kramarz (2001), Moment estimation with attrition: An application to economic models, *Journal of the American Statistical Association*, 96, 1223–1231.

Abrevaya, J., and S.G. Donald (2010), A GMM approach for dealing with missing data on regressors and instruments, Manuscript.

Allen, M.R., P. Gillett, A. Kettleborough, G. Hegerl, R. Schnur, A. Stott, G. Boer, C. Covey, L. Delworth, S. Jones, B. Mitchell, and T.P. Barnett (2006), Quantifying anthropogenic influence on recent near-surface temperature change, *Surveys in Geophysics*, 27, 491–544.

Anderson, T.L., J. Charlson, E. Schwartz, R. Knutti, O. Boucher, H. Rodhe, and J. Heintzenberg (2003), Climate forcing by aerosols–a hazy picture, *Science*, 300, 1103–1104.

Andreae, M.O., D. Jones, and P.M. Cox (2005), Strong present-day aerosol cooling implies a hot future, *Nature*, 435, 1187–1190.

Angrist, J., V. Lavy, and A. Schlosser (2006), Multiple experiments for the causal link between the quantity and quality of children, Working Paper No. 06-26, Department of Economics, Massachusetts Institute of Technology.

Anselin, L. (1988), *Spatial Econometrics: Methods and Models*, Kluwer, Dordrecht.

Anselin, L., and A.K. Bera (1998), Spatial dependence in linear regression models with an introduction to spatial econometrics, in: A. Ullah and D.E.A. Giles, Eds., *Handbook of Applied Economic Statistics*, Marcel Dekker, New York, 237–289.

Arellano, M. (2003), *Panel Data Econometrics*, Oxford University Press, Oxford.

Arellano, M., and S. Bond (1991), Some tests of specification for panel data: Monte carlo evidence and an application to employment equations, *The Review of Economic Studies*, 58, 277–297.

Arrow, K.J. (1971), *Essays in the Theory of Risk Bearing*, North-Holland, Amsterdam.

Arrow, K.J. (1974), The use of unbounded utility functions in expected-utility maximization: Response, *Quarterly Journal of Economics*, 88, 136–138.

Böhringer, C., A. Löschel, and T.F. Rutherford (2007), Decomposing the integrated assessment of climate change, *Journal of Economic Dynamics and Control*, 31, 683–702.

Baltagi, B.H. (2001), *Econometric Analysis of Panel Data*, Wiley, Chichester.

Baltagi, B.H., and B. Raj (1992), A survey of recent theoretical developments in the econometrics of panel data, *Empirical Economics*, 17, 85–109.

Baltagi, B.H., and J.M. Griffin (1988), A generalized error component model with heteroskedastic disturbances, *International Economic Review*, 39, 745–753.

Baltagi, B.H., S.H. Song, and W. Koh (2003), Testing panel data regression models with spatial error correlation, *Journal of Econometrics*, 117, 123–150.

Baltagi, B.H., S.H. Song, B.C. Jung, and W. Koh (2007), Testing for serial correlation, spatial autocorrelation and random effects using panel data, *Journal of Econometrics*, 140, 5–51.

Barillas, F., and J. Fernández-Villaverde (2007), A generalization of the endogenous grid method, *Journal of Economic Dynamics and Control*, 31, 2698–2712.

Barr, J.R., and A.S. Manne (1967), Numerical experiments with finite horizon planning models, *Indian Economic Review*, 2, 1–29.

Barro, R.J. (2009), Rare disasters, asset prices, and welfare costs, *American Economic Review*, 99, 243–264.

Barro, R.J., and J.F. Ursúa (2008), Consumption disasters in the twentieth century, *American Economic Review*, 98, 58–63.

Bellouin, N., O. Boucher, J. Haywood, and M.S. Reddy (2005), Global estimate of aerosol direct radiative forcing from satellite measurements, *Nature*, 438, 1138–1141.

Bickel, P., C. Klaassen, Y. Ritov, and J. Wellner (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins University Press, Baltimore.

Binswanger, H.P. (1980), Attitude toward risk: Experimental measurement in rural India, *American Journal of Agricultural Economics*, 62, 395–407.

Blundell, R., and S. Bond (1998), Initial conditions and moment restrictions in dynamic panel data models, *Journal of Econometrics*, 87, 115–143.

Bobenrieth, E.S.A., J.R.A. Bobenrieth, and B.D. Wright (2008), A foundation for the solution of consumption-saving behavior with a borrowing constraint and unbounded marginal utility, *Journal of Economic Dynamics and Control*, 32, 695–708.

Budyko, M.I. (1969), The effect of solar radiation variations on the climate of the Earth, *Tellus XXI*, 5, 611–619.

Burr, I.W. (1942), Cumulative frequency functions, *Annals of Mathematical Statistics*, 13, 215–232.

Burr, I.W., and P.J. Cislak (1968), On a general system of distributions: I Its curve-shape characteristics, II The sample median, *Journal of the American Statistical Association*, 63, 627–635.

Chamberlain, G. (1987), Asymptotic efficiency in estimation with conditional moment restrictions, *Journal of Econometrics*, 34, 305–334.

Chen, B., G. Yi, and R. Cook (2010), Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random, *Journal of the American Statistical Association*, 105, 336–353.

Chen, X., H. Hong, and A. Tarozzi (2008), Semiparametric efficiency in GMM models with auxiliary data, *Annals of Statistics*, 36, 808–843.

Chiappori, P.-A., and M. Paiella (2008), Relative risk aversion is constant: Evidence from panel data, Discussion Paper No. 5/2008, Department of Economic Studies, University of Naples "Parthenope", Italy.

Chichilnisky, G. (2000), An axiomatic approach to choice under uncertainty with catastrophic risks, *Resource and Energy Economics*, 22, 221–231.

Christiano, L.J., and J.D.M. Fisher (2000), Algorithms for solving dynamic models with occasionally binding constraints, *Journal of Economic Dynamics and Control*, 24, 1179–1232.

Crutzen, P.J., and V. Ramanathan (2003), The parasol effect on climate, *Science*, 302, 1678–1680.

Dasgupta, P., and K.-G. Mäler (2003), The economics of non-convex ecosystems: Introduction, *Environmental and Resource Economics*, 26, 499–525.

Denuit, M., and L. Eeckhoudt (2010), Stronger measures of higher-order risk attitudes, *Journal of Economic Theory*, 145, 2027–2036.

Dorofeenko, V., G.S. Lee, and K.D. Salyer (2010), A new algorithm for solving dynamic stochastic macroeconomic models, *Journal of Economic Dynamics and Control*, 34, 388–403.

Doroodian, K., and R. Boyd (2003), The linkage between oil price shocks and economic growth with inflation in the presence of technological advances: A CGE model, *Energy Policy*, 31, 989–1006.

Driscoll, J.C., and A.C. Kraay (1998), Consistent covariance matrix estimation with spatially dependent panel data, *The Review of Economics and Statistics*, 80, 549–560.

Elashoff, M., and L. Ryan (2004), An EM algorithm for estimating equations, *Journal of Computational and Graphical Statistics*, 13, 48–65.

Eyckmans, J., and H. Tulkens (2003), Simulating coalitionally stable burden sharing agreements for the climate change problem, *Resource and Energy Economics*, 25, 299–327.

Feng, H., and J. Zhao (2006), Alternative intertemporal permit trading regimes with stochastic abatement costs, *Resource and Energy Economics*, 28, 24–40.

Finetti, B. de (1952), Sulla preferibilità, *Giornale degli Economisti e Annali di Economia*, 11, 685–709.

Fishburn, P.C. (1976), Unbounded utility functions in expected utility theory, *Quarterly Journal of Economics*, 90, 163–168.

Friend, I., and M.E. Blume (1975), The demand for risky assets, *American Economic Review*, 65, 900–922.

Gerber, H.U. (1979), *An Introduction to Mathematical Risk Theory*, S.S. Huebner Foundation Monograph 8, Irwin, Homewood, IL.

Geweke, J. (2001), A note on some limitations of CRRA utility, *Economics Letters*, 71, 341–345.

Gilgen, H., and A. Ohmura (1999), The global energy balance archive, *Bulletin of the American Meteorological Society*, 80, 831–850.

Gollier, C. (2001), *The Economics of Risk and Time*, MIT Press, Cambridge, MA.

Gollier, C. (2002), Time horizon and the discount rate, *Journal of Economic Theory*, 107, 463–473.

Gollier, C. (2008), Discounting with fat-tailed economic growth, *Journal of Risk and Uncertainty*, 37, 171–186.

Graham, B. (2010), Efficiency bounds for missing data models with semiparametric restrictions, *Econometrica, forthcoming*.

Gregory, J.M., J. Stouffer, B. Raper, A. Stott, and N.A. Rayner (2002), An observationally based estimate of the climate sensitivity, *Journal of Climate*, 15, 3117–3121.

Harrison, G.W., J.A. List, and C. Towe (2007), Naturally occurring preferences and exogenous laboratory experiments: A case study of risk aversion, *Econometrica*, 75, 433–458.

Haywood, J., and O. Boucher (2000), Estimates of the direct and indirect radiative forcings due to tropospheric aerosols: A review, *Reviews of Geophysics*, 38, 513–543.

Heyde, C., and R. Morton (1996), Quasi-likelihood and generalizing the EM algorithm, *Journal of the Royal Statistical Society. Series B*, 58, 317–327.

Hirano, K., G. Imbens, and G. Ridder (2003), Efficient estimation of average treatment effects using the estimated propensity score, *Econometrica*, 71, 1161–1189.

Hoel, M., and L. Karp (2002), Taxes versus quotas for a stock pollutant, *Resource and Energy Economics*, 24, 367–384.

Holt, C.A., and S.K. Laury (2002), Risk aversion and incentive effects, *American Economic Review*, 92, 1644–1655.

Ikefuji, M., R.J.A. Laeven, J.R. Magnus, and C. Muris (2010), Catastrophe, VSL, and curvature, Technical Notes, Tilburg University.

IPCC (2007): Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, and H.L. Miller, Eds., *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Johnson, N.L., S. Kotz, and N. Balakrishnan (1995), *Continuous Univariate Distributions*, Vol. 2, 2nd ed., John Wiley, New York.

Judd, K.L. (1992), Projection methods for solving aggregate growth models, *Journal of Economic Theory*, 58, 410–452.

Köbberling, V., and P.P. Wakker (2005), An index of loss aversion, *Journal of Economic Theory*, 122, 119–131.

Kaplow, L. (2005), The value of a statistical life and the coefficient of relative risk aversion, *Journal of Risk and Uncertainty*, 31, 23–34.

Kapoor, M., H.H. Kelejian, and I.R. Prucha (2007), Panel data models with spatially correlated error components, *Journal of Econometrics*, 140, 97–130.

Kaufman, Y.J., D. Tanré, and O. Boucher (2002), A satellite view of aerosols in the climate system, *Nature*, 419, 215–223.

Keller, K., B.M. Bolker, and D.F. Bradford (2004), Uncertain climate thresholds and optimal economic growth, *Journal of Environmental Economics and Management*, 48, 723–741.

Kelly, D.L., and C.D. Kolstad (1999), Bayesian learning, growth, and pollution, *Journal of Economic Dynamics and Control*, 23, 491–518.

Krawczyk, J.B. (2005), Coupled constraint Nash equilibria in environmental games, *Resource and Energy Economics*, 27, 157–181.

Lau, M.I., A. Pahlke, and T.F. Rutherford (2002), Approximating infinite-horizon models in a complementarity format: A primer in dynamic general equilibrium analysis, *Journal of Economic Dynamics and Control*, 26, 577–609.

Leach, A.J. (2007), The climate change learning curve, *Journal of Economic Dynamics and Control*, 31, 1728–1752.

Leach, A.J. (2009), The welfare implications of climate change policy, *Journal of Environmental Economics and Management*, 57, 151–165.

Leandri, M. (2009), The shadow price of assimilative capacity in optimal flow pollution control, *Ecological Economics*, 68, 1020–1031.

Li, Q., and T. Stengos (1994), Adaptive estimation in the panel data error component model with heteroskedasticity of unknown form, *International Economic Review*, 35, 981–1000.

Little, R. J. A., and D. B. Rubin (2002), *Statistical analysis with missing data*, Wiley Series in Probability and Statistics, New York.

Magnus, J.R. (1982), Multivariate error component analysis of linear and nonlinear regression models by maximum likelihood, *Journal of Econometrics*, 19, 239–285.

Magnus, J.R. (1988), *Linear Structures*, Griffin's Statistical Monographs and Courses, No. 42, Edward Arnold, London and Oxford University Press, New York.

Magnus, J.R., and H. Neudecker (1988), *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley and Sons, Chichester/New York. Second edition, 1999.

Manne, A.S., and R.G. Richels (1992), *Buying Greenhouse Insurance: The Economic Costs of CO2 Emission Limits*, MIT Press, Cambridge, Mass.

Mastrandrea, M.D., and S.H. Schneider (2004), Probabilistic integrated assessment of 'dangerous' climate change, *Science*, 304, 571–575.

McCormick, M.P., W. Thomason, and C.R. Trepte (1995), Atmospheric effects of the Mt Pinatubo eruption, *Nature*, 373, 399–404.

McGuffie, K., and A. Henderson-Sellers (2001), Forty years of numerical climate modelling, *International Journal of Climatology*, 21, 1067–1109.

Menger, K. (1934), Das Unsicherheitsmoment in der Wertlehre, *Zeitschrift für Nationalökonomie*, 5, 459–485.

Mitchell, T.D., and P.D. Jones (2005), An improved method of constructing a database of monthly climate observations and associated high-resolution grids, *International Journal of Climatology*, 25, 693–712.

Montgomery, W.D. (1972), Markets in licenses and efficient pollution control programs, *Journal of Economic Theory*, 5, 395–418.

Moscadelli, M. (2004), The modelling of operational risk: Experience with the analysis of the data collected by the Basel Committee, Technical Report 517, Banca d'Italia.

Myhre, G. (2009), Consistency between satellite-derived and modeled estimates of the direct aerosol effect, *Science*, 325, 187–190.

Neumann, J. von, and O. Morgenstern (1944, 1947, 1953), *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ.

Newey, W. (1990), Semiparametric efficiency bounds, *Journal of Applied Econometrics*, 5, 99–135.

Newey, W. (2001), Conditional moment restrictions in censored and truncated regression models, *Econometric Theory*, 17, 863–888.

Newey, W., and D. McFadden (1994), Large sample estimation and hypothesis testing, *Handbook of Econometrics*, 4, 2111–2245.

Nordhaus, W.D. (1994), *Managing the Global Commons: The Economics of Climate Change*, MIT Press, Cambridge, Mass.

Nordhaus, W.D. (2008), *A Question of Balance: Weighing the Options on Global Warming Policies*, Yale University Press, New Haven, CT.

Nordhaus, W.D. (2009), An analysis of the dismal theorem, Cowles Foundation Discussion Paper 1686, Yale University.

Nordhaus, W.D., and Z. Yang (1996), A regional dynamic general-equilibrium model of alternative climate-change strategies, *American Economic Review*, 86, 741–765.

Norris, J.R., and M. Wild (2007), Trends in aerosol radiative effects over Europe inferred from observed cloud cover, solar 'dimming,' and solar 'brightening', *Journal of Geophysical Research*, 112, D08214.

North, G.R. (1975), Theory of energy-balance climate models, *Journal of the Atmospheric Sciences*, 32, 2033–2042.

North, G.R., F. Cahalan, and J.A. Coakley Jr. (1981), Energy balance climate models, *Reviews of Geophysics and Space Physics*, 19, 91–121.

Post, T., M.J. van den Assem, G. Baltussen, and R.H. Thaler (2008), Deal or no deal? Decision making under risk in a large-payoff game show, *American Economic Review*, 98, 38–71.

Power, H.C. (2003), Trends in solar radiation over Germany and an assessment of the role of aerosols and sunshine duration, *Theoretical and Applied Climatology*, 76, 47–63.

Pratt, J.W. (1964), Risk aversion in the small and in the large, *Econometrica*, 32, 122–136.

Räisänen, J. (2007), How reliable are climate models?, *Tellus A*, 59, 2–29.

Rabin, M. (2000), Risk aversion and expected-utility theory: A calibration theorem, *Econometrica*, 68, 1281–1292.

Ramanathan, V., C. Chung, D. Kim, T. Bettge, L. Buja, T. Kiehl, M. Washington, Q. Fu, R. Sikka, and M. Wild (2005), Atmospheric brown clouds: Impacts on South Asian climate and hydrological cycle, *Proceedings of the National Academy of Sciences*, 102, 5326–5333.

Ramanathan, V., J. Crutzen, T. Kiehl, and D. Rosenfeld (2001), Aerosols, climate, and the hydrological cycle, *Science*, 294, 2119–2124.

Ranjan, R., and J. Shortle (2007), The environmental Kuznets curve when the environment exhibits hysteresis, *Ecological Economics*, 64, 204–215.

Robins, J. M., and A. Rotnitzky (1995), Semiparametric efficiency in multivariate regression models with missing data, *Journal of the American Statistical Association*, 90, 122–129.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994), Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association*, 89, 846–866.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995), Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association*, 90, 106–121.

Robinson, P.M. (1988), Root-N consistent semiparametric regression, *Econometrica*, 56, 931–954.

Roe, G.H., and M.B. Baker (2007), Why is climate sensitivity so unpredictable?, *Science*, 318, 629–632.

Roughgarden, T., and S.H. Schneider (1999), Climate change policy: Quantifying uncertainties for damages and optimal carbon taxes, *Energy Policy*, 27, 415–429.

Ryan, T.M. (1974), The use of unbounded utility functions in expected-utility maximization: Comment, *Quarterly Journal of Economics*, 88, 133–135.

Savage, L.J. (1954), *The Foundations of Statistics*, Wiley, New York.

Schwartz, S.E. (2007), Heat capacity, time constant, and sensitivity of Earth's climate system, *Journal of Geophysical Research*, 112, D24S05.

Searle, S.R., and H.V. Henderson (1979), Dispersion matrices for variance components models, *Journal of the American Statistical Association*, 74, 465–470.

Sellers, W.D. (1969), A global climatic model based on the energy balance of the Earth-atmosphere system, *Journal of Applied Meteorology*, 8, 392–400.

Severini, T., and G. Tripathi (2001), A simplified approach to computing efficiency bounds in semiparametric models, *Journal of Econometrics*, 102, 23–66.

Sims, C.A. (2001), Pitfalls of a minimax approach to model uncertainty, *American Economic Review*, 91, 51–54.

Skiba, A.K. (1978), Optimal growth with a convex-concave production function, *Econometrica*, 46, 527–540.

Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, B. Averyt, M. Tignor, and H.L. Miller, Eds. (2007), *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, United Kingdom.

Stacy, E.W. (1962), A generalization of the gamma distribution, *Annals of Mathematical Statistics*, 33, 1187–1192.

Stern, N. (2007), *The Economics of Climate Change: The Stern Review*, Cambridge University Press, Cambridge, UK.

Stott, P.A., B. Mitchell, R. Allen, L. Delworth, M. Gregory, A. Meehl, and B.D. Santer (2006), Observational constraints on past attributable warming and predictions of future global warming, *Journal of Climate*, 19, 3055–3069.

Subbotin, M.Th. (1923), On the law of frequency of error, Mathematicheskii Sbornik, 31, 296–301.

Sugden, R. (1997), Alternatives to expected utility theory: Foundations and concepts, in: *Handbook of Expected Utility Theory*, Eds. S. Barberà, P.J. Hammond and C. Seidl, Kluwer, Boston, Mass.

Tett, S.F.B., S. Jones, A. Stott, C. Hill, B. Mitchell, R. Allen, J. Ingram, C. Johns, E. Johnson, A. Jones, L. Roberts, H. Sexton, and M.J. Woodage (2002), Estimation of natural and anthropogenic contributions to twentieth century temperature change, *Journal of Geophysical Research*, 107, 4306.

Trenberth, K.E., T. Fasullo, and J. Kiehl (2009), Earth's global energy budget, *Bulletin of the American Meteorological Society*, 90, 311–323.

Vaart, A. van der (2000), *Asymptotic statistics*, Cambridge University Press, Cambridge.

Verbeek, M., and Nijman, T. (1992), Testing for selectivity bias in panel data models, *International Economic Review*, 33, 681–703.

Wakker, P.P. (2008), Explaining the characteristics of the power (CRRA) utility family, *Health Economics*, 17, 1329–1344.

Wansbeek, T., and A. Kapteyn (1982), A class of decompositions of the variance-covariance matrix of a generalized error components model, *Econometrica*, 50, 713–724.

Weitzman, M.L. (2009), On modeling and interpreting the economics of catastrophic climate change, *Review of Economics and Statistics*, 91, 1–19.

Wild, M. (2009), Global dimming and brightening: A review, *Journal of Geophysical Research*, 114, D00D16.

Wirl, F. (1991), The political economics of Wackersdorf: Why do politicians stick to their past decisions?, *Public Choice*, 70, 343–350.

Wooldridge, J. (2007), Inverse probability weighted estimation for general missing data problems, *Journal of Econometrics*, 141, 1281–1301.

Yaari, M. (1969), Some remarks on measures of risk aversion and their uses, *Journal of Economic Theory*, 1, 315–329.

Zeeuw, A. de (2008), Dynamic effects on the stability of international environmental agreements, *Journal of Environmental Economics and Management*, 55, 163–174.